Distribution and Symmetric Distribution Regression Model for Histogram-Valued Variables

Sónia Dias[1,⋆], **Paula Brito**[2]

1. Escola Superior de Tecnologia e Gestão do Instituto Politécnico Viana do Castelo & LIAAD-INESC Porto LA, Universidade do Porto, Portugal
2. Faculdade de Economia & LIAAD-INESC Porto LA, Universidade do Porto, Portugal
⋆Contact author: sdias@estg.ipvc.pt

Abstract

Histogram-valued variables are a particular kind of variables studied in *Symbolic Data Analysis* where to each entity under analysis corresponds a distribution that may be represented by a histogram or by a quantile function. Linear regression models for this type of data are necessarily more complex than a simple generalization of the classical model: the parameters cannot be negative still the linear relationship between the variables must be allowed to be either direct or inverse. In this work we propose a new linear regression model for histogram-valued variables that solves this problem, named *Distribution and Symmetric Distribution Regression Model*. To determine the parameters of this model it is necessary to solve a quadratic optimization problem, subject to non-negativity constraints on the unknowns; the error measure between the predicted and observed distributions uses the Mallows distance. As in classical analysis, the model is associated with a goodness-of-fit measure whose values range between 0 and 1. Using the proposed model, applications with real and simulated data are presented.

Keywords: data with variability; linear regression; Symbolic Data Analysis; quantile functions; Mallows distance.

1 Introduction

Classical multivariate statistics studies data tables that summarize observations made on "statistical units" (individuals); each row of the table represents one individual and each of these individuals is characterized by different variables (in columns). The "values" attained by the variables may be real values if the variable represents the measurement of a quantity (quantitative variables) or a category if the variable is qualitative. As an example, let us have classical quantitative variables such as the age, weight and height of a particular football player. The observations of these data are typically represented in classical data tables, but how can we represent the result of the weight of the football player if we don't know his exact weight? And what if we are interested in studying the age, weight and height not of one single player but of a football team? In the first situation, the individuals are described by attributes whose associated values are quantitative values that cannot be "measured" with precision. In cases like this, we are in the presence of imprecise data. In the second situation we are interested in describing one class of individuals.

The "best values" attained by the variables that characterize each class are not real values or categories but sets of "values", intervals or distributions. Even though data with variability or uncertainty may be represented by the same type of elements, the meaning of these elements is different. For example, the interval $[80, 82]$ may mean that the weight of one football player is between 80 and 82 Kg. On the other hand, the interval $[75, 80]$ may represent the weights of all players from a given football team. In the first situation the interval represents the imprecision of the weight value, whereas in the second situation the interval considers the variability of weight values in the football team.

In this research we will focus on situations where variability in data description occurs. The classical solution to analyze these data is to reduce the collection of records associated to each individual or class of individuals to one value, this may be the mean, mode or maximum/minimum; however, with this option the variability across the records is lost. In alternative to the classical analysis applied to these kind of data, Diday [11] introduced *Symbolic Data Analysis*, where the term *symbolic data* refers precisely to data with variability. To understand the concept of symbolic data it is important to assess where variability comes from. The variability of the data might emerge due to the aggregation of observations [2] that can be contemporary, if the records are collected in the same temporal instant or the temporal instant is not relevant, and temporal if the time is the aggregation criterion and the records are grouped along one unit of time, for example one day. In both situations, the initial data or micro-data, are organized in classical data tables where each individual, termed first-level unit, is described by classical variables. Depending on the type of aggregation, the construction of the symbolic data table is different. When the aggregation is temporal, the entities under analysis are the original first-level units, now characterized by sets of values originating from the records collected over a unit of time. In situations where the aggregation is contemporary, the entities - higher-level units - are classes of individuals (sets of first-level units) grouped according to specific characteristics. In this situation, the variables describing both the higher-level and the respective first-level units are the same; however the "values" that the variables take for each higher-level unit are now sets of values or functions obtained from the respective first-level units.

Following the definition of Bock and Diday [7], a symbolic variable Y is a mapping $Y : E \to \mathbb{B}$ defined on a set E of statistical entities ($E = \Omega = \{1, 2, \ldots, m\}$ when the individuals are first-level units or $E = \{C_1, C_2, \ldots\}$ with $C_j \subseteq \Omega$ when the individuals are higher-level units) and which takes its values in a set $\mathbb{B}$. Henceforth in this work, when we use the term unit, we will be referring to a first-level unit or to a higher-level unit, according to the kind of prior aggregation of the micro-data used to build the symbolic data table.

Similarly to the classical case, symbolic variables can also be classified as quantitative or qualitative, according to the nature of the elements of $\mathbb{B}$. For quantitative symbolic variables, each unit is allowed to take a single value (single-valued variables); a finite set of values (multi-valued variables); an interval (interval-valued variables); or a mapping that can be a probability/frequency/weight distribution (modal-valued variables).

In this paper, we will be dealing with a particular type of modal-valued variables, the *histogram-valued variables*. In this case, the values attained by the variable for each unit are empirical frequency distributions or, more specifically, histograms, where the values in each subinterval are assumed to be uniformly

distributed. If we consider a symbolic variable where all units are associated to one only interval of real numbers (uniformly distributed) with probability/frequency/weight equal to one, then we are in the presence of *interval-valued variables*.

As an example, consider a symbolic data table containing information about patients (adults) attending healthcare centers, during a fixed period of time. In healthcare centre A, the age of patients ranged from 25 to 53 years old, in healthcare centre B, it ranged from 33 to 68 years old and in healthcare centre C, the age of patients ranged from 20 to 75 years old, so that the age is an interval-valued variable. Now consider another variable which records the waiting time for consultations. In this case, information is recorded for 5 time lengths (0 to 15 minutes, 15 to 30 minutes,...), and the corresponding symbolic variable is therefore a histogram-valued variable (see Table 1). Notice that in this example the entities under analysis are the healthcare centers (higher-level units), for each of which we have aggregated information (contemporary aggregation), and NOT the individual patients attending each centre (first-level units).

Healthcare centers	Age	Waiting Time (minutes)
A	$[25, 53]$	$\{[0, 15[\,, 0; [15, 30[\,, 0.25; [30, 45[\,, 0.5; [45, 60[\,, 0; \geq 60, 0.25\}$
B	$[33, 68]$	$\{[0, 15[\,, 0.25; [15, 30[\,, 0.25; [30, 45[\,, 0.25; [45, 60[\,, 0.25; \geq 60, 0\}$
C	$[20, 75]$	$\{[0, 15[\,, 0.33; [15, 30[\,, 0; [30, 45[\,, 0.33; [45, 60[\,, 0; \geq 60, 0.33\}$

Table 1: Data for three healthcare centers.

Symbolic Data Analysis has achieved considerable development since the eighties of last century (see, for instance, [3], [4], [7], [12], [19]). Recently, there has been a growing interest in the analysis of histogram-valued variables, though still more research is developed for interval-valued variables. The methods proposed so far for the former are indeed, frequently, a generalization of their counterparts for the latter. The main definitions of descriptive statistics for one, two or more histogram-valued variables have already been studied. Billard and Diday [4] defined mean, observed and relative frequency, empirical density function, empirical joint density function; for variance and covariance two definitions were proposed [3]; [4]; [5]; Irpino and Verde [14] defined distribution functions and joint distribution functions.

The first definitions and methods for histogram-valued variables are generally obtained from the application of the classic concepts to the midpoints of the histograms' subintervals, using the respective weights. Furthermore, although the symbolic variables' values are distributions and not real numbers, the results of the application of these concepts are real numbers. For example, the mean of m observations of the histogram-valued variable, proposed by Billard and Diday [4], is a real number. It should be noticed, however, that in recent years other works have been put forward where the "results" are already distributions. For example, Irpino and Verde [14] present an alternative definition of mean for histogram-valued variables, which produces a mean distribution, that they termed by *barycentric histogram*.

Work with histogram-valued variables has been recently reported in different domains, such as Principal Component Analysis [21], [22]; Cluster Analysis [14]; Time series [1] and Linear Regression [5], [23].

The first linear regression model for histogram-valued variables was a generalization of the first model

proposed for interval-valued variables by Billard and Diday [3], [6]. Other models have also been proposed for interval-valued variables [17], [18]; however, these models present some limitations: firstly, they are based on differences between real values and do not appropriately quantify the closeness between intervals; then, the elements predicted by the models may fail to build an interval; the most recent model imposes non-negativity constraints on the coefficients, therefore forcing a direct linear relationship. These limitations prevent a generalization of the models to histogram-valued variables, so that alternative models are being developed (see, e.g., [13], [23]). Our goal is to propose a linear regression model for histogram-valued variables allowing predicting distributions from other distributions, without forcing a direct linear relationship.

The development of non-descriptive methods for Symbolic Data Analysis is still an open research topic for almost all kinds of symbolic variables. Notice, however, papers recently published proposing probabilistic models for interval-valued variables [8],[16].

The remaining of the paper is organized as follows. Section 2 introduces histogram-valued variables in more detail, and presents a short study about the space of the quantile functions. In Section 3, the problem of defining a linear regression model for histogram-valued variables is addressed. A model and a respective goodness-of-fit measure are proposed. Section 4 reports results of a simulation study and two examples that illustrate the application of the model. Finally, Section 5 concludes the paper, pointing out directions for future research.

2 Symbolic Data Analysis: histogram data

2.1 Histogram-valued variables

Consider a symbolic variable $Y : E \to \mathbb{B}$. The set of units E may be $E = \Omega = \{1, 2, \ldots, m\}$ when the individuals are first-level units or $E = \{C_1, C_2, \ldots\}$ with $C_j \subseteq \Omega$ when the individuals are higher-level units. Consider also the quantitative (single-value) variable $\dot{Y}$ defined on a set Ω. If the aggregation of the observations is temporal, to each unit $j \in \Omega$ corresponds the empirical distribution of the values that $\dot{Y}$ takes within a certain unit of time. If the aggregation is contemporary, to each unit j corresponds the empirical distribution of $\dot{Y}$ in C_j. As histograms are a usual representation of empirical distributions, this kind of symbolic variables are termed *histogram-valued variables*. More generally we can define histogram-valued variables as follows:

Definition 2.1 *Y is a histogram-valued variable when to each unit j corresponds a empirical distribution $Y(j)$, that can be represented by a histogram [7], [4]:*

$$H_{Y(j)} = \left\{ \left[\underline{I}_{Y(j)_1}, \overline{I}_{Y(j)_1} \right[, p_{j1}; \left[\underline{I}_{Y(j)_2}, \overline{I}_{Y(j)_2} \right[, p_{j2}; \ldots; \left[\underline{I}_{Y(j)n_j}, \overline{I}_{Y(j)n_j} \right] , p_{jn_j} \right\} \tag{1}$$

where $\underline{I}_{Y(j)_i}$ and $\overline{I}_{Y(j)_i}$ represent the lower and upper bound of the interval i; p_{ji} is the frequency associated to the subinterval $\left[\underline{I}_{Y(j)_i}, \overline{I}_{Y(j)_i} \right[$ with $i \in \{1, 2, \ldots, n_j\}$, n_j is the number of subintervals

for the j^{th} unit, $j = 1, \ldots, m$ $\sum_{i=1}^{n_j} p_{ij} = 1$, $\underline{I}_{Y(j)_i} \leq \overline{I}_{Y(j)_i}$ *and* $\overline{I}_{Y(j)_i} \leq \underline{I}_{Y(j)_{i+1}}$.

Alternatively, $Y(j)$ can be represented by the inverse of the cumulative empirical distribution function, also called quantile function $\Psi^{-1}_{Y(j)}$ [14]:

$$\Psi^{-1}_{Y(j)}(t) = \begin{cases} \underline{I}_{Y(j)_1} + \frac{t}{w_{j1}} a_{Y(j)_1} & if \quad 0 \leq t < w_{j1} \\ \underline{I}_{Y(j)_2} + \frac{t - w_{j1}}{w_{j2} - w_{j1}} a_{Y(j)_2} & if \quad w_{j1} \leq t < w_{j2} \\ \vdots & \\ \underline{I}_{Y(j)_{n_j}} + \frac{t - w_{jn_j-1}}{1 - w_{jn_j-1}} a_{Y(j)_{n_j}} & if \quad w_{jn_j-1} \leq t \leq 1 \end{cases} \tag{2}$$

where $w_{jl} = \begin{cases} 0 & if \quad l = 0 \\ \sum_{h=1}^{l} p_{jh} & if \quad l = 1, \ldots, n_j \end{cases}$ *and* $a_{Y(j)_i} = \overline{I}_{Y(j)_i} - \underline{I}_{Y(j)_i}$ *with* $i \in \{1, \ldots, n_j\}$;

n_j is the number of subintervals in $Y(j)$.

Or, considering the subintervals of the histograms defined by the centers $c_{Y(j)_i}$ and half-ranges $r_{Y(j)_i}$, the representation of the $Y(j)$ can be given by

$$H_{Y(j)} = \left\{ \left[c_{Y(j)_1} - r_{Y(j)_1}, c_{Y(j)_1} + r_{Y(j)_1} \right[, p_{j1}; \ldots; \left[c_{Y(j)n_j} - r_{Y(j)n_j}, c_{Y(j)n_j} + r_{Y(j)n_j} \right], p_{jn_j} \right\} \tag{3}$$

or

$$\Psi^{-1}_{Y(j)}(t) = \begin{cases} c_{Y(j)_1} + \left(2\frac{t}{w_{j1}} - 1 \right) r_{Y(j)_1} & if \quad 0 \leq t < w_{j1} \\ c_{Y(j)_2} + \left(2\frac{t - w_{j1}}{w_{j2} - w_{j1}} - 1 \right) r_{Y(j)_2} & if \quad w_{j1} \leq t < w_{j2} \\ \vdots & \\ c_{Y(j)_{n_j}} + \left(2\frac{t - w_{jn_j-1}}{1 - w_{jn_j-1}} - 1 \right) r_{Y(j)_{n_j}} & if \quad w_{jn_j-1} \leq t \leq 1 \end{cases} \tag{4}$$

Any of these representations of the empirical distribution that each unit takes can be termed histogram value. Henceforth, when we use the term distribution, we are referring to an empirical distribution of a continuous variable. Furthermore, it is also assumed that within each subinterval $\left[\underline{I}_{Y(j)_i}, \overline{I}_{Y(j)_i} \right[$ the values for the variable Y for each unit $j = 1, \ldots, m$, are uniformly distributed.

If any of the weights p_{ji} with $i > 1$ is nullo, the function $\Psi_{Y(j)}$ doesn't have inverse with domain between 0 and 1. Consequently the function $\Psi^{-1}_{Y(j)}$ is not continuous and has $n_j - 1$ pieces. In this case it is not

possible to calculate the value of $\Psi^{-1}_{Y(j)}(w_{ji-1})$ *but only* $\lim_{t\to w^-_{ji-1}} \Psi^{-1}_{Y(j)}(t)$ *and* $\lim_{t\to w^+_{ji-1}} \Psi^{-1}_{Y(j)}(t)$.

When $n_j = 1$ and for each unit j, $Y(j)$ takes values only on the interval $\left[\underline{I}_{Y(j)}, \overline{I}_{Y(j)}\right[$ with frequency $p_j = 1$, the histogram-valued variable is then reduced to the particular case of an interval-valued variable. In this case, the quantile function is given by

$$\Psi^{-1}_{Y(j)}(t) = \underline{I}_{Y(j)} + \left(\overline{I}_{Y(j)} - \underline{I}_{Y(j)}\right) t, \quad with \quad 0 \leq t \leq 1. \tag{5}$$

When we work with histogram-valued variables, it is important to note that for different observations, the number of subintervals in the histograms or the pieces in functions may be different; the subintervals of histograms $H_{Y(j)}$ are considered ordered and disjoint, and if this is not the case, it must be possible to rewrite them in the required form [26], [2].

Example 2.1 *Consider the histograms*

$$H_X = \{[1, 3[\,, 0.1; [3, 5[\,, 0.6; [5, 8]\,, 0.3\}$$

and

$$H_Y = \{[0, 1[\,, 0.8; [1, 4]\,, 0.2\}$$

that caracterize an unit for the histogram-valued variables X *and* Y*, respectively. These histograms are represented in Figure 1:*

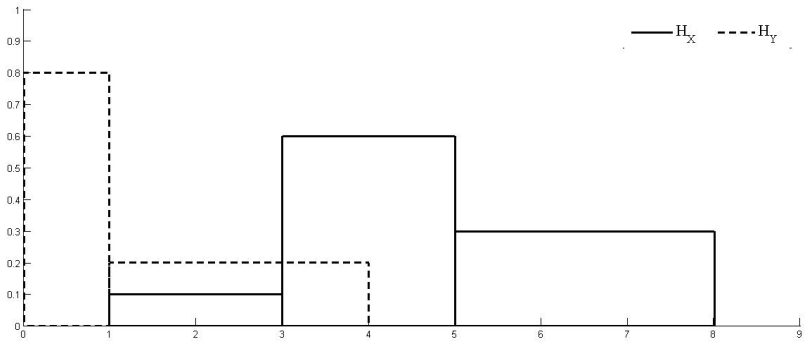

Figure 1: Representation of the histograms H_X and H_Y in Example 2.1.

Alternatively, these histograms can be represented by their quantile functions (see Figure 2):

$$\Psi^{-1}_X(t) = \begin{cases} 1 + \frac{t}{0.1} \times 2 & if \quad 0 \leq t < 0.1 \\ 3 + \frac{t-0.1}{0.6} \times 2 & if \quad 0.1 \leq t < 0.7 \\ 5 + \frac{t-0.7}{0.3} \times 3 & if \quad 0.7 \leq t \leq 1 \end{cases} \qquad \Psi^{-1}_Y(t) = \begin{cases} \frac{t}{0.8} & if \quad 0 \leq t < 0.8 \\ 1 + \frac{t-0.8}{0.2} \times 3 & if \quad 0.8 \leq t \leq 1 \end{cases}$$

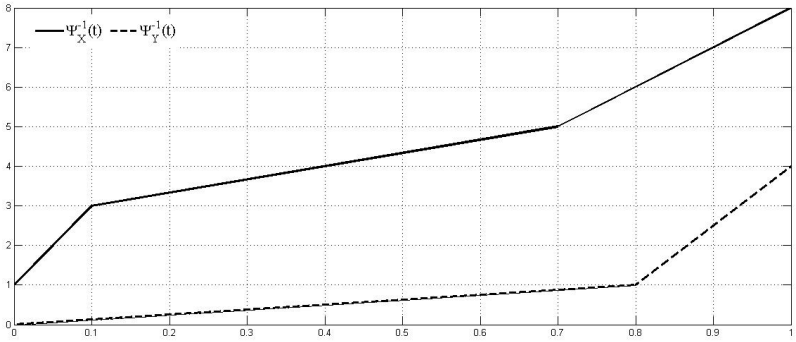

Figure 2: Representation of the quantile functions Ψ_X^{-1} and Ψ_Y^{-1} in Example 2.1.

It is important to bear in mind that in a histogram the lower bound of each subinterval is always less than or equal to the upper bound, $\underline{I}_{Y(j)i} \leq \overline{I}_{Y(j)i}$ and the upper bound of the following subinterval is always greater or equal to the previous, $\overline{I}_{Y(j)i} \leq \underline{I}_{Y(j)i+1}$. Consequently, the quantile function that represents the empirical distribution is always a non-decreasing function in the domain $[0, 1]$.

Many concepts and methods for histogram-valued variables have been defined using the representation of their realizations in the form of histograms [3], [4]. Only in more recent studies have these variables' values been represented as quantile functions [1], [14], [24], [25]. When the distributions are represented as histograms, the choice of the arithmetic becomes crucial. The complexity of the arithmetics [9], [26] that have been proposed so far for histograms was arguably the reason why the distributions started being represented as quantile functions. If we represent the distribution that each unit takes on a histogram-valued variable by a quantile function, then operations are simplified because, as quantile functions are piecewise functions, the adequate arithmetic for them is a function arithmetic. In this work the option is to represent the distributions by quantile functions. However, this representation raises other questions.

To operate with quantile functions, it is necessary to define all functions involved with an equal number of pieces or, equivalently, to rewrite all correspondent histograms with the same number of subintervals. For this, it may be necessary to apply the procedure defined by Irpino and Verde [14]. In addition, it is important to avoid that the number of subintervals for each histogram becomes "too" large (which could happen by applying the process proposed by Irpino and Verde [14]), in which case the distributions that represent the data would be meaningless. To prevent this situation, we may consider the suggestion of Colombo [9] who encountered similar problems, and has considered advantageous to work with equiprobable histograms (histograms of equal probability subintervals).

2.2 The space of quantile functions

Quantile functions are a particular kind of functions. If we consider the set of the functions defined from $\mathbb{R}$ in $\mathbb{R}$, $\mathcal{F}(\mathbb{R}, \mathbb{R})$ and the usual operations defined in $\mathcal{F}$: addition $(f + g)(x) = f(x) + g(x), \forall x \in \mathbb{R}$ and

product of a function by a real number $(\lambda f)(x) = \lambda f(x), \forall x \in \mathbb{R}$, and $\lambda \in \mathbb{R}$, it follows that $(\mathcal{F}, +, .)$ is a vector space. However, if we consider the particular case of the set of the quantile functions, $\mathcal{E}([0,1], \mathbb{R})$, defined on $[0,1]$, we don't have a subspace of the vector space $(\mathcal{F}, +, .)$. Analyzing the behavior of these operations it is possible to understand why $\mathcal{E}([0,1], \mathbb{R})$, with the usual operations, does not verify the vector space definition.

Consider the quantile functions $\Psi_X^{-1}(t)$ and $\Psi_Y^{-1}(t)$ defined according to (2) in *Definition 2.1* both with n subintervals, after having been rewritten in accordance with the process described in [14]. These functions represent the distributions that the histogram-valued variables X and Y take for one unit. The addition of these quantile functions leads to the function

$$\Psi_X^{-1}(t) + \Psi_Y^{-1}(t) = \begin{cases} \underline{I}_{X_1} + \underline{I}_{Y_1} + \frac{t}{w_1}(a_{X_1} + a_{Y_1}) & if \quad 0 \leq t < w_1 \\ \underline{I}_{X_2} + \underline{I}_{Y_2} + \frac{t-w_1}{w_2-w_1}(a_{X_2} + a_{Y_2}) & if \quad w_1 \leq t < w_2 \\ \vdots & \\ \underline{I}_{X_n} + \underline{I}_{Y_n} + \frac{t-w_{n-1}}{1-w_{n-1}}(a_{X_n} + a_{Y_n}) & if \quad w_{n-1} \leq t \leq 1 \end{cases}$$

When we add two quantile functions we obtain a non-decreasing function. In this case both the slope and the y-intercept of the resulting function are influenced by the two functions.

The particular case of the addition of a quantile function $\Psi_X^{-1}(t)$ with a real number α is the function

$$\left(\Psi_X^{-1} + \alpha\right)(t) = \begin{cases} \underline{I}_{X_1} + \alpha + \frac{t}{w_1} a_{X_1} & if \quad 0 \leq t < w_1 \\ \underline{I}_{X_2} + \alpha + \frac{t-w_1}{w_2-w_1} a_{X_2} & if \quad w_1 \leq t < w_2 \\ \vdots & \\ \underline{I}_{X_n} + \alpha + \frac{t-w_{n-1}}{1-w_{n-1}} a_{X_n} & if \quad w_{n-1} \leq t \leq 1 \end{cases}$$

In this case, only the y-intercept is affected by the operation, we have a translation up when adding a real positive number α and a translation down when the real number α is negative.

The multiplication of the quantile function $\Psi_X^{-1}(t)$ by a real number λ leads to the function

$$\lambda\Psi_X^{-1}(t) = \begin{cases} \lambda\underline{I}_{X_1} + \frac{t}{w_1}(\lambda a_{X_1}) & if \quad 0 \leq t < w_1 \\ \lambda\underline{I}_{X_2} + \frac{t-w_1}{w_2-w_1}(\lambda a_{X_2}) & if \quad w_1 \leq t < w_2 \\ \vdots & \\ \lambda\underline{I}_{X_n} + \frac{t-w_{n-1}}{1-w_{n-1}}(\lambda a_{X_n}) & if \quad w_{n-1} \leq t \leq 1 \end{cases}$$

In this case, both the slope and the y-intercept are affected by λ. If λ is positive we will have a non-

decreasing function but if λ is negative we will obtain a decreasing function that cannot be a quantile function, because quantile functions must always be non-decreasing functions. It is for this reason that the $\mathcal{E}([0,1],\mathbb{R})$, is a semi-vectorial space.

The following example illustrates this situation.

Example 2.2 *Consider the distribution represented by the quantile function $\Psi_X^{-1}(t)$ presented in Example 2.1. If we multiply the quantile function $\Psi_X^{-1}(t)$ by the positive real number 2, we obtain a non-decreasing function but if we multiply the quantile function $\Psi_X^{-1}(t)$ by the negative real number -1 the resulting function is not a non-decreasing function. The following functions and representations in Figure 3 illustrate this situation.*

$$2\Psi_X^{-1}(t) = \begin{cases} 2 + \frac{t}{0.1} \times 4 & if \quad 0 \leq t < 0.1 \\ 6 + \frac{t-0.1}{0.6} \times 4 & if \quad 0.1 \leq t < 0.7 \\ 10 + \frac{t-0.7}{0.3} \times 6 & if \quad 0.7 \leq t \leq 1 \end{cases}$$

$$-\Psi_X^{-1}(t) = \begin{cases} -1 + \frac{t}{0.1} \times (-2) & if \quad 0 \leq t < 0.1 \\ -3 + \frac{t-0.1}{0.6} \times (-2) & if \quad 0.1 \leq t < 0.7 \\ -5 + \frac{t-0.7}{0.3} \times (-3) & if \quad 0.7 \leq t \leq 1 \end{cases}$$

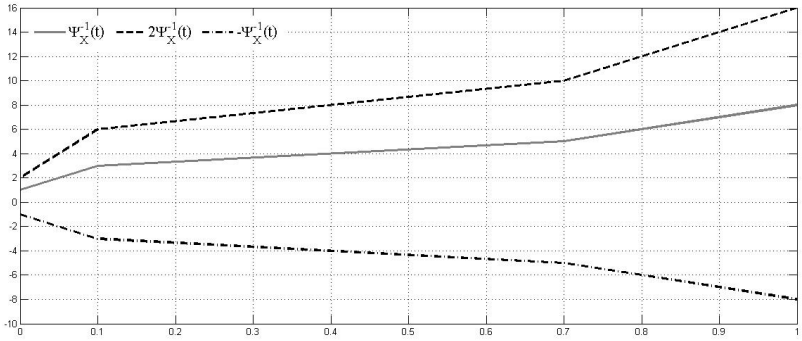

Figure 3: Representation of the functions $\Psi_X^{-1}(t)$, $2\Psi_X^{-1}(t)$, $-\Psi_X^{-1}(t)$ in Example 2.2.

In conclusion, $\mathcal{E}([0,1],\mathbb{R})$, is not a vector space because the elements of this space do not have symmetric elements. If we have a quantile function $\Psi_X(t)$, the function $-\Psi_X^{-1}(t)$ is not a non-decreasing function and consequently cannot be a quantile function. However if we consider the distributions represented by histograms and use the histograms arithmetic proposed by Colombo [9] it is possible to obtain a new histogram, that is the symmetric of the histogram H_X. The histogram $-H_X$ is the symmetric of the histogram H_X if $-H_X$ and H_X are symmetric in relation to the $yy-$axis.

As an example of the situation above, *Figure 4* represents the histogram H_X in Example 2.1 and the respective symmetric histogram.

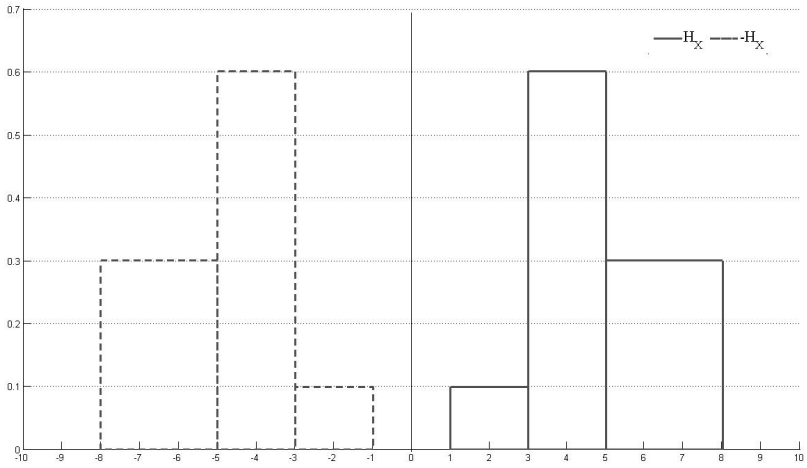

Figure 4: Representation of the histogram H_X in Example 2.1 and the respective symmetric histogram $-H_X$.

It is obviously possible to define the quantile function that represents the distribution of the histogram $-H_X$. This quantile function is $-\Psi_X^{-1}(1-t)$ with $t \in [0,1]$ and is not the function obtained by multiplying the quantile function $\Psi_X^{-1}(t)$ by -1. *Figure 5* shows that the function $-\Psi_X^{-1}(t)$ in Example 2.2 is different from the quantile function $-\Psi_X^{-1}(1-t)$ that corresponds to the histogram $-H_X$.

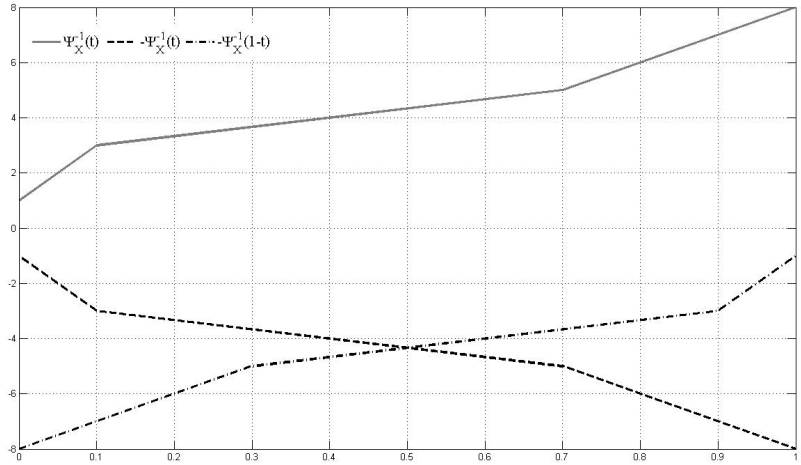

Figure 5: Representation of the functions $\Psi_X^{-1}(t)$, $-\Psi_X^{-1}(t)$, and $-\Psi_X^{-1}(1-t)$, in Example 2.2.

To conclude this section, it is important to underline some conclusions about the function $-\Psi_X^{-1}(1-t)$, $t \in [0,1]$:

- As it is required for quantile functions, $-\Psi_X^{-1}(1-t)$ is a non-decreasing function;
- $\Psi_X^{-1}(t) - \Psi_X^{-1}(1-t)$ is not a null function, as expected, but is a quantile function with null (symbolic) mean [4];
- the functions $-\Psi_X^{-1}(1-t)$ and $\Psi_X^{-1}(t)$ are linearly independent, providing that $-\Psi_X^{-1}(1-t) \neq \Psi_X^{-1}(t)$;
- $-\Psi_X^{-1}(1-t) = \Psi_X^{-1}(t)$ only when the histogram H_X is symmetric with respect to the $yy-$axis

3 Linear Regression Model for histogram-valued variables

The first linear regression model for histogram-valued variables was proposed by Billard and Diday [5]. This model is a generalization of the *Center Model* [6] defined by the same authors for interval-valued variables but, with this model it is possible that the predicted results are not histogram values. Because of this, recently same studies have emerged in an attempt to find new proposals for a linear regression model for this kind of variables. A recent model has been proposed by Verde and Irpino [23].

Our main goal in this work is to propose a linear regression model for histogram-valued variables. More precisely, to provide a linear regression model that considers data with variability and allows predicting histogram values.

To define this model, three problems need to be solved:

- Find an error measure to quantify the difference between the observed and predicted distributions represented by histograms or quantile functions;
- Define a linear regression model for histogram-valued variables that allows predicting histograms or their quantile functions from other histograms or quantile functions, without forcing a direct linear relationship;
- Measure the goodness-of-fit of the model.

3.1 Error measure

In classical linear regression, to quantify the error between the observed values y_j and the predicted values $\widehat{y}_j$ the difference between two real numbers, $e_j = y_j - \widehat{y}_j$ is used. In this case, the model to estimate the values $\widehat{y}_j$ minimizes the quantity $\sum_{j=1}^{m}(y_j - \widehat{y}_j)^2$. However, due to the complexity of histogram-valued variables, the error between the observed and predicted distributions requires a different approach.

In their work about forecasting time series, applied to histogram-valued variables, Arroyo and Maté [1], [2] also needed to measure the error between the observed and forecasted distributions. Therefore, they sought for a good measure to analyze the similarity between two distributions. Firstly, they considered the possibility of computing the difference between two distributions represented by their respective histograms using the histograms' arithmetic. However, this option turned out to be of little use. As we have seen, it is not easy to operate with the histograms arithmetic and some results are not as expected. This shows that it is not adequate to analyze the similarity between distributions with this concept. The options of those authors were to use dissimilarity measures for distributions and they opted for the Wasserstein and Mallows distance [15], [1] to measure the diference between the observed and forecasted distributions. The justification for the choice of the Wasserstein and Mallows distance was the fact that they are distances and thus present interesting properties for error measurement: positive definiteness, symmetry, and triangle inequality condition. On the other hand, for Arroyo and Maté [1], [2], the Mallows distance is the one that better adjusts to the concept of distance as assessed by the human eye. This distance was also used in other works such as Irpino and Verde [14], where the Mallows distance is used to determine the *barycentric histogram* and is then successfully applied to cluster histogram data. The same authors used this distance in their linear regression model for histogram-valued variables [23].

In using the Wasserstein and Mallows distances, the distributions taken by the histogram-valued variables are represented by their quantile functions. These distances are defined as follows:

Definition 3.1 *Given two quantile functions $\Psi^{-1}_{X(j)}(t)$ and $\Psi^{-1}_{Y(j)}(t)$ that represent the distributions that the histogram-valued variables X and Y take at unit j, the Wasserstein distance is defined as:*

$$D_W(\Psi^{-1}_{X(j)}(t), \Psi^{-1}_{Y(j)}(t)) = \int_0^1 \left|\Psi^{-1}_{X(j)}(t) - \Psi^{-1}_{Y(j)}(t)\right| dt \tag{6}$$

and the Mallows distance:

$$D_M(\Psi^{-1}_{X(j)}(t), \Psi^{-1}_{Y(j)}(t)) = \sqrt{\int_0^1 (\Psi^{-1}_{X(j)}(t) - \Psi^{-1}_{Y(j)}(t))^2 dt} \tag{7}$$

Instead of using the quantile functions that represent the distributions, Irpino and Verde [14] rewrote the Mallows distance using the histograms, more specifically the centre and half-range of their subintervals. The square of the Mallows distance can be also defined as follows:

Property 3.1 *Consider two histogram-valued variables X and Y. The distributions that these variables take for a given unit j, can be represented by the quantile functions $\Psi^{-1}_{X(j)}(t)$ and $\Psi^{-1}_{Y(j)}(t)$ or the histograms $H_{X(j)}$ and $H_{Y(j)}$. The square of the Mallows distance between these distributions is given by*

$$D^2_M(\Psi^{-1}_{X(j)}(t), \Psi^{-1}_{Y(j)}(t)) = \sum_{i=1}^{n} p_i \left[(c_{X(j)_i} - c_{Y(j)_i})^2 + \frac{1}{3}(r_{X(j)_i} - r_{Y(j)_i})^2\right]$$

where, relatively to the histogram-valued variables X or Y for unit j :

- $c_{X(j)_i} = \frac{\overline{I}_{X(j)_i} + \underline{I}_{X(j)_i}}{2}$ *and* $c_{Y(j)_i} = \frac{\overline{I}_{Y(j)_i} + \underline{I}_{Y(j)_i}}{2}$ *are the centers of the intervals* i*, with* $i \in \{1, \ldots, n\}$;
- $r_{X(j)_i} = \frac{\overline{I}_{X(j)_i} - \underline{I}_{X(j)_i}}{2}$ *and* $r_{Y(j)_i} = \frac{\overline{I}_{Y(j)_i} - \underline{I}_{Y(j)_i}}{2}$ *are the half-ranges of the intervals* i*, with* $i \in \{1, \ldots, n\}$.

It seems therefore appropriate to choose the Wasserstein or Mallows distance to measure the similarity between the observed and predicted distributions by the linear regression model. Because of the properties of the absolute value function we choose to define the error measure between two distributions with the Mallows distance.

Definition 3.2 *Consider, for each unit* j*,* $\Psi^{-1}_{Y(j)}(t)$ *the quantile function of the observed distribution* $Y(j)$ *and* $\Psi^{-1}_{\widehat{Y}(j)}(t)$ *the quantile function that represents the predicted distribution* $\widehat{Y}(j)$*. The error between* $Y(j)$ *and* $\widehat{Y}(j)$ *is defined by:*

$$SE(j) = D_M^2(\Psi^{-1}_{Y(j)}(t), \Psi^{-1}_{\widehat{Y}(j)}(t)) \tag{8}$$

The total error is the sum of the errors, that according to Property 3.1, may be written as follows:

$$SE = \sum_{j=1}^{m} D_M^2(\Psi^{-1}_{Y(j)}(t), \Psi^{-1}_{\widehat{Y}(j)}(t)) = \sum_{j=1}^{m}\sum_{i=1}^{n} p_{ji}\left[(c_{Y(j)_i} - c_{\widehat{Y}(j)_i})^2 + \frac{1}{3}(r_{Y(j)_i} - r_{\widehat{Y}(j)_i})^2\right] \tag{9}$$

3.2 The *DSD Regression Model*

The first option to define the functional linear relation between histogram data was to adapt the classical model to these data. Consider that we want to predict the distributions that the histogram-valued variable Y takes from p histogram-valued variables X_k with $k \in \{1, \ldots, p\}$. At unit j, $j \in \{1, \ldots, m\}$, the predicted distribution $\widehat{Y}(j)$ would than be obtained as follows:

$$\widehat{Y}(j) = \gamma + \alpha_1 X_1(j) + \alpha_2 X_2(j) + \ldots + \alpha_p X_p(j).$$

As already mentioned, in this work we choose to represent the distributions by quantile functions. However, when we multiply a quantile function by a negative number we do not obtain a non-decreasing function. Therefore, it is necessary to impose positivity restrictions on the parameters of the model. Denoting by $\Psi^{-1}_{\widehat{Y}(j)}(t)$ the quantile function of the predicted distribution $\widehat{Y}(j)$, we obtain the linear regression model as follows:

$$\Psi^{-1}_{\widehat{Y}(j)}(t) = \beta_0 + \beta_1 \Psi^{-1}_{X_1(j)}(t) + \beta_2 \Psi^{-1}_{X_2(j)}(t) + \ldots + \beta_p \Psi^{-1}_{X_p(j)}(t)$$

with $\beta_k \geq 0$ and $k \in \{1, 2, \ldots, p\}$.

The non-negativity constraints imposed on the coefficients force a direct linear relationship, and limitations similar to those present in linear regression models defined for interval-valued variables occur (see, e.g., [17]). Although we did not generalize the model for interval-valued variables to histogram-valued variables, in defining a model that allows to predict a quantile function from other quantile functions, we obtain a model with the same limitations as observed before.

It is not possible to have negative parameters in the previous model. Neverthless, it is fundamental to allow for the possibility of a direct and an inverse linear relation between the variable Y and the variables X_k. For this reason, our proposal is to include in the linear regression model both the quantile functions $\Psi^{-1}_{X_k(j)}(t)$, that represent the distributions that the histogram-valued variables X_k take for each unit j, and the quantile functions that represent the respectively symmetric histograms $-\Psi^{-1}_{X_k(j)}(1-t)$ (see also *Section 2.2*). Therefore we proposed the following model:

Definition 3.3 *Consider the histogram-valued variables $X_1; X_2; \ldots; X_p$. The quantile functions that represent the distribution that these histogram-valued variables take for each unit j are $\Psi^{-1}_{X_1(j)}(t)$, $\Psi^{-1}_{X_2(j)}(t), \ldots, \Psi^{-1}_{X_p(j)}(t)$ and the quantile functions that represent the respective symmetric histograms associated to each unit of the refered variables are $-\Psi^{-1}_{X_1(j)}(1-t), -\Psi^{-1}_{X_2(j)}(1-t), \ldots, -\Psi^{-1}_{X_p(j)}(1-t)$, with $t \in [0,1]$. Each quantile function $\Psi^{-1}_{Y(j)}$, can be expressed as follows:*

$$\Psi^{-1}_{Y(j)}(t) = \Psi^{-1}_{\widehat{Y}(j)}(t) + \varepsilon_j(t).$$

where $\Psi^{-1}_{\widehat{Y}(j)}(t)$ is the predicted quantile function for unit j, obtained from

$$\begin{aligned}\Psi^{-1}_{\widehat{Y}(j)}(t) &= \gamma + \alpha_1\Psi^{-1}_{X_1(j)}(t) - \beta_1\Psi^{-1}_{X_1(j)}(1-t) + \alpha_2\Psi^{-1}_{X_2(j)}(t) - \beta_2\Psi^{-1}_{X_2(j)}(1-t) + \\ &\quad + \ldots + \alpha_p\Psi^{-1}_{X_p(j)}(t) + \beta_p\Psi^{-1}_{X_p(j)}(1-t).\end{aligned}$$

with $t \in [0,1]$; $\alpha_k, \beta_k \geq 0$, $k \in \{1, 2, \ldots, p\}$ and $\gamma \in \mathbb{R}$.

The error, for each unit j, is the piecewise function given by $\varepsilon_j(t) = \Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)$.

For each unit j, the predicted distribution $\widehat{Y}(j)$ can be represented by the quantile function $\Psi^{-1}_{\widehat{Y}(j)}$ or by the respective histogram $H_{\widehat{Y}(j)}$. This linear regression model will be named ***Distribution and Symmetric Distribution (DSD) Regression Model****.*

Consider the particular case of the linear regression model where there is only one explicative histogram-valued variable X. In this case we can obtain the quantile function $\Psi^{-1}_{Y(j)}(t)$, for each unit j, by the model:

$$\Psi^{-1}_{Y(j)}(t) = \gamma + \alpha\Psi^{-1}_{X(j)}(t) - \beta\Psi^{-1}_{X(j)}(1-t) + \varepsilon_j(t) \tag{10}$$

with $\alpha, \beta \geq 0$, and $\gamma \in \mathbb{R}$.

When including in the model both the distribution of the explicative histogram-valued variables, and the respective symmetric distributions, the restrictions on the parameters are imposed; however, this does not imply a direct linear relationship. In the particular case of (10), we consider that the linear regression is direct if $\alpha > \beta$ and inverse if $\alpha < \beta$.

3.3 Parameters of the *DSD Regression Model*

In classical statistics, the parameters of the linear regression model are estimated solving the minimization problem $\sum_{j=1}^{m}(y_j - \widehat{y}_j)^2$, where y_j are the observed and $\widehat{y}_j$ the predicted values, respectively, with $j \in \{1, \ldots, m\}$. To solve this problem the least squares method is used.

For histogram-valued variables the parameters of the *DSD Model*, in *Definition 3.3*, are estimated solving a quadratic optimization problem, subject to non-negativity constraints on the unknowns.

Definition 3.4 *Consider $\Psi^{-1}_{\widehat{Y}(j)}(t)$ obtained by the DSD Model. The quadratic optimization problem is written as:*

$$Minimize \qquad SE = \sum_{j=1}^{m} D_M^2(\Psi^{-1}_{Y(j)}(t), \Psi^{-1}_{\widehat{Y}(j)}(t))$$

with $\alpha_k, \beta_k \geq 0$, $k \in \{1, 2, \ldots, p\}$ and $\gamma \in \mathbb{R}$.

To present more specifically the function to minimize, it is important to define all the quantile functions involved in this expression considering the conditions referred in *Section 2.1*. The quantile functions that represent the distributions taken by X_k and the respective symmetric, for a given unit j are, respectively:

$$\Psi^{-1}_{X_k(j)}(t) = \begin{cases} c_{X_k(j)_1} + \left(2\frac{t}{w_1} - 1\right) r_{X_k(j)_1} & if \quad 0 \leq t < w_1 \\ c_{X_k(j)_2} + \left(2\frac{t-w_1}{w_2-w_1} - 1\right) r_{X_k(j)_2} & if \quad w_1 \leq t < w_2 \\ \vdots & \\ c_{X_k(j)_n} + \left(2\frac{t-w_{(n-1)}}{1-w_{(n-1)}} - 1\right) r_{X_k(j)_n} & if \quad w_{n-1} \leq t \leq 1 \end{cases} \tag{11}$$

$$-\Psi^{-1}_{X_k(j)}(1-t) = \begin{cases} -c_{X_k(j)_n} + \left(2\frac{t}{w_1} - 1\right) r_{X_k(j)_n} & if \quad 0 \leq t < w_1 \\ -c_{X_k(j)_{n-1}} + \left(2\frac{t-w_1}{w_2-w_1} - 1\right) r_{X_k(j)_{n-1}} & if \quad w_1 \leq t < w_2 \\ \vdots & \\ -c_{X_k(j)_1} + \left(2\frac{t-w_{n-1}}{1-w_{n-1}} - 1\right) r_{X_k(j)_1} & if \quad w_{n-1} \leq t \leq 1 \end{cases} \tag{12}$$

According to the *DSD Model*, the quantile function that represents the distribution taken by the predicted histogram-valued variable $\widehat{Y}$, for a given unit j is:

$$\Psi^{-1}_{\widehat{Y}(j)}(t) = \begin{cases} \sum_{k=1}^{p}\left(\alpha_k c_{X_k(j)_1} - \beta_k c_{X_k(j)_n}\right) + \gamma + \left(2\frac{t}{w_1} - 1\right)\sum_{k=1}^{p}\left(\alpha_k r_{X_k(j)_1} + \beta_k r_{X_k(j)_n}\right) & if \quad 0 \le t < w_1 \\ \sum_{k=1}^{p}\left(\alpha_k c_{X_k(j)_2} - \beta_k c_{X_k(j)_{n-1}}\right) + \gamma + \left(2\frac{t-w_1}{w_2-w_1} - 1\right)\sum_{k=1}^{p}\left(\alpha_k r_{X_k(j)_2} + \beta_k r_{X_k(j)_{n-1}}\right) & if \quad w_1 \le t < w_2 \\ \vdots & \\ \sum_{k=1}^{p}\left(\alpha_k c_{X_k(j)_n} - \beta_k c_{X_k(j)_1}\right) + \gamma + \left(2\frac{t-w_{n-1}}{1-w_{n-1}} - 1\right)\sum_{k=1}^{p}\left(\alpha_k r_{X_k(j)_n} + \beta_k r_{X_k(j)_1}\right) & if \quad w_{n-1} \le t \le 1 \end{cases} \tag{13}$$

Similarly, for unit j, the quantile function that represents the distribution taken by the histogram-valued variable, Y is

$$\Psi^{-1}_{Y(j)}(t) = \begin{cases} c_{Y(j)_1} + \left(2\frac{t}{w_1} - 1\right) r_{Y(j)_1} & if \quad 0 \le t < w_1 \\ c_{Y(j)_2} + \left(2\frac{t-w_1}{w_2-w_1}\right) r_{Y(j)_2} & if \quad w_1 \le t < w_2 \\ \vdots & \\ c_{Y(j)_n} + \left(2\frac{t-w_{n-1}}{1-w_{n-1}}\right) r_{Y(j)_n} & if \quad w_{n-1} \le t \le 1 \end{cases} \tag{14}$$

Consider these quantile functions and the Mallows distance defined according to *Property 3.1*. The quadratic optimization problem presented in *Definition 3.4* can then be rewritten as follows:

$$\begin{aligned} Minimize \quad SE \quad = & \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left[\left(c_{Y(j)_i} - \sum_{k=1}^{p}\left(\alpha_k c_{X_k(j)_i} - \beta_k c_{X_k(j)_{n-i+1}}\right) - \gamma\right)^2 \right. \\ & \left. + \tfrac{1}{3}\left(r_{Y(j)_i} - \sum_{k=1}^{p}\left(\alpha_k r_{X_k(j)_i} + \beta_k r_{X_k(j)_{n-i+1}}\right)\right)^2\right] \end{aligned} \tag{15}$$

$$\text{subject to } \alpha_k, \beta_k \ge 0,\ k \in \{1, 2, \ldots, p\} \text{ and } \gamma \in \mathbb{R}.$$

Or, in matricial form:

$$Minimize \quad SE = \frac{1}{2}B^T H B + F^T B + C \tag{16}$$

$$\text{subject to } -\alpha_k, -\beta_k \le 0;\ k \in \{1, 2, \ldots, p\} \text{ and } \gamma \in \mathbb{R}.$$

In this latter case, $H = [h_{lq}]$ is the hessian matrix, a symmetric matrix of order $2p+1$, with p the number of variables X_k. The elements of the symmetric matrix H are defined as follows:

$$h_{lq} = \begin{cases} \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(2c_{X_{\frac{l+1}{2}(j)_i}} c_{X_{\frac{q+1}{2}(j)_i}} + \frac{2}{3} r_{X_{\frac{l+1}{2}(j)_i}} r_{X_{\frac{q+1}{2}(j)_i}} \right) & if \quad \textit{l,q are odd and } l, q \leq 2p \\ \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(2c_{X_{\frac{l}{2}(j)_{n-i+1}}} c_{X_{\frac{q}{2}(j)_{n-i+1}}} + \frac{2}{3} r_{X_{\frac{l}{2}(j)_{n-i+1}}} r_{X_{\frac{q}{2}(j)_{n-i+1}}} \right) & if \quad \textit{l,q are even and } l, q \leq 2p \\ \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(-2c_{X_{\frac{l}{2}(j)_{n-i+1}}} c_{X_{\frac{q+1}{2}(j)_i}} + \frac{2}{3} r_{X_{\frac{l}{2}(j)_{n-i+1}}} r_{X_{\frac{q+1}{2}(j)_i}} \right) & if \quad \textit{l is even, q is odd and } l, q \leq 2p \\ \sum_{j=1}^{m}\sum_{i=1}^{n} 2p_i c_{X_{\frac{q+1}{2}(j)_i}} & if \quad \textit{q is odd and } l = 2p+1 \\ \sum_{j=1}^{m}\sum_{i=1}^{n} -2p_i c_{X_{\frac{q}{2}(j)_{n-i+1}}} & if \quad \textit{q is even and } l = 2p+1 \end{cases}$$

The vector column of independent terms, $F = [f_l]$ with $2p+1$ rows is given by:

$$f_l = \begin{cases} \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(-2c_{Y(j)_i} c_{X_{\frac{l+1}{2}(j)_i}} - \frac{2}{3} r_{Y(j)_i} r_{X_{\frac{l+1}{2}(j)_i}} \right) & if \quad \textit{l is odd and } l \leq 2p \\ \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(2c_{Y(j)_i} c_{X_{\frac{l+1}{2}(j)_{n-i+1}}} - \frac{2}{3} r_{Y(j)_i} r_{X_{\frac{l+1}{2}(j)_{n-i+1}}} \right) & if \quad \textit{l is even and } l \leq 2p \\ \sum_{j=1}^{m}\sum_{i=1}^{n} -2p_i c_{Y(j)_i} & if \quad l = 2p+1 \end{cases}$$

The elements of the matrices H and F are computed from the first order partial derivatives of the function SE in (15). These derivatives are presented in *Appendix A*. Finally, the vector column of the parameters, B, and the real value C, are defined as follows:

$$B = [\alpha_1 \quad \beta_1 \quad \alpha_2 \quad \beta_2 \quad \dots \quad \alpha_p \quad \beta_p \quad \gamma]^T$$

and

$$C = \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(c^2_{Y(j)_i} + \frac{1}{3} r^2_{Y(j)_i} \right).$$

For each particular situation, it is possible to solve this quadratic optimization problem, subject to non-negativity on the constraints, and find the optimal solution. Consider the optimal solution for this optimization problem,

$$B^* = [\alpha_1^* \quad \beta_1^* \quad \alpha_2^* \quad \beta_2^* \quad \cdots \quad \alpha_n^* \quad \beta_n^* \quad \gamma^*]^T.$$

Afterwards, it is possible to predict the distributions $\widehat{Y}(j)$, for each $j \in \{1, \dots, m\}$, considering the obtained matrix B^*. Each predicted distribution may be represented by the quantile function as in (13) or by the respective histogram

$$H_{\widehat{Y}(j)} = \left\{ \left[\sum_{k=1}^{p} \left(\alpha_k^* \underline{I}_{X_k(j)_1} - \beta_k^* \overline{I}_{X_k(j)_n} \right) + \gamma^*, \sum_{k=1}^{p} \left(\alpha_k^* \overline{I}_{X_k(j)_1} - \beta_k^* \underline{I}_{X_k(j)_n} \right) + \gamma^* \right], p_1; \ \ldots \right.$$
$$\left. \ldots; \left[\sum_{k=1}^{p} \left(\alpha_k^* \underline{I}_{X_k(j)_n} - \beta_k^* \overline{I}_{X_k(j)_1} \right) + \gamma^*, \sum_{k=1}^{p} \left(\alpha_k^* \overline{I}_{X_k(j)_n} - \beta_k^* \underline{I}_{X_k(j)_1} \right) + \gamma^* \right], p_n \right\}$$

Consider the minimization problem defined in (15) or matricially in (16). The optimal solution of the quadratic optimization problem, subject to non-negativity constraints, verifies the Kuhn Tucker conditions [27]. Therefore, the optimal solution B^* for this optimization problem, for all $k \in \{1, \ldots, p\}$ verifies the following conditions:

- $-\alpha_k^*, -\beta_k^* \leq 0$;
- $\frac{\partial SE(B^*)}{\partial \alpha_k} \geq 0$; $\frac{\partial SE(B^*)}{\partial \beta_k} \geq 0$; $\frac{\partial SE(B^*)}{\partial \gamma} = 0$; $\frac{\partial SE(B^*)}{\partial \alpha_k} \alpha_k^* = 0$; $\frac{\partial SE(B^*)}{\partial \beta_k} \beta_k^* = 0$;

From the Kuhn Tucker conditions, it is possible to prove some properties associated with the predicted distribution. Some of these are the counterparts of the corresponding properties in classical statistics, and will allow defining a measure to evaluate the goodness-of-fit of the model. Before describing these properties, it is necessary to present two important definitions of the concept of mean for histogram-valued variables.

Definition 3.5 *[4] Consider the histogram-valued variable Y. For each unit j, with $j \in \{1, \ldots, m\}$, $Y(j)$ may be represented by the histogram defined in (4). The mean of variable Y is defined as follows:*

$$\overline{Y} = \frac{1}{m} \sum_{j=1}^{m} \left(\sum_{i=1}^{n_j} c_{Y(j)_i} p_{ji} \right).$$

where n_j is the number of subintervals for the j^{th} unit.

Irpino and Verde [14] defined the *barycentric histogram* as the histogram that is at a minimum distance - in the sense of the Mallows distance - of the m distributions. In this case, a mean distribution is obtained instead of a mean that is a real number.

The quantile function of the *barycentric histogram* is the same as the mean quantile function, that is computed from the average of the m quantile functions that represent the m given distributions. The mean quantile function is defined as follows:

Definition 3.6 *Consider the m quantile functions $\Psi_{Y(j)}^{-1}(t)$, $j \in \{1, \ldots, m\}$, all defined with n pieces. The mean quantile function $\overline{\Psi_Y^{-1}}(t)$ is the function where each piece is the mean of the corresponding m pieces involved. The function is then,*

$$\overline{\Psi_Y^{-1}}(t) = \begin{cases} \sum_{j=1}^{m} \frac{c_{Y(j)_1}}{m} + \left(2\frac{t}{w_1} - 1\right)\frac{r_{Y(j)_1}}{m} & if \quad 0 \le t < w_1 \\ \sum_{j=1}^{m} \frac{c_{Y(j)_2}}{m} + \left(2\frac{t-w_1}{w_2-w_1} - 1\right)\frac{r_{Y_{j2}}}{m} & if \quad w_1 \le t < w_2 \\ \vdots & \\ \sum_{j=1}^{m} \frac{c_{Y(j)_n}}{m} + \left(2\frac{t-w_{n-1}}{1-w_{n-1}} - 1\right)\frac{r_{Y_{jn}}}{m} & if \quad w_{n-1} \le t \le 1 \end{cases}$$

So, we have $\overline{\Psi_Y^{-1}}(t) = \frac{1}{m}\sum_{j=1}^{m} \Psi_{Y(j)}^{-1}(t)$.

These two concepts of mean for histogram-valued variables are related as we can see in the following proposition.

Proposition 3.1 *Considering the mean quantile function* $\overline{\Psi_Y^{-1}}(t)$ *of the histogram-valued variable* Y *and its mean* $\overline{Y}$*, we have*

$$\overline{Y} = \int_0^1 \overline{\Psi_Y^{-1}}(t)dt.$$

This result is due to Irpino and Verde [24] and may easily be proved considering *Definitions 3.5* and *3.6*.

Now, considering the previous results and the Kuhn Tucker conditions, we may prove the following properties.

Property 3.2 *For each unit* j*, let* $\widehat{Y}(j)$ *be the distribution predicted by the DSD Model and consider the parameters obtained for the optimal solution* $B^* = [\alpha_1^* \quad \beta_1^* \quad \alpha_2^* \quad \beta_2^* \quad \cdots \quad \alpha_n^* \quad \beta_n^* \quad \gamma^*]^T$*. The mean of the predicted histogram-valued variable* $\overline{\widehat{Y}}$ *is given by:*

$$\overline{\widehat{Y}} = \sum_{k=1}^{p} (\alpha_k^* - \beta_k^*)\,\overline{X_k} + \gamma^*.$$

Proof: Each observation j, of the predicted histogram-valued variable $\widehat{Y}(j)$, can be represented by the quantile function as in (13) considering for parameters the optimal solution B^*, of the quadratic optimization problem in (15). As such, the mean quantile function $\Psi_{\widehat{Y}}^{-1}$ can be calculated by *Definition 3.6*. So, applying *Proposition 3.1* we can prove that $\overline{\widehat{Y}} = \sum_{k=1}^{p} (\alpha_k^* - \beta_k^*)\,\overline{X_k} + \gamma^*$. □

Property 3.3 *The mean of the predicted histogram-valued variable* $\overline{\widehat{Y}}$ *is equal to the mean of the observed histogram-valued variable* $\overline{Y}$*.*

Proof: Consider the function to minimize in (15),

$$SE = \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left[\left(c_{Y(j)_i} - \sum_{k=1}^{p} \left(\alpha_k c_{X_{k(j)i}} - \beta_k c_{X_{k(j)n-i+1}} \right) - \gamma \right)^2 + \frac{1}{3} \left(r_{Y(j)_i} - \sum_{k=1}^{p} \left(\alpha_k r_{X_{k(j)i}} + \beta_k r_{X_{k(j)n-i+1}} \right) \right)^2 \right].$$

For the optimal solution B^* we have $\frac{\partial SE(B^*)}{\partial \gamma} = 0$. Consequently,

$$\begin{aligned}
&2\sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(\sum_{k=1}^{p} \alpha_k^* c_{X_{k(j)i}} \right) - 2\sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(\sum_{k=1}^{p} \beta_k^* c_{X_{k(j)(n-i+1)}} \right) + 2m\gamma^* - 2\sum_{j=1}^{m}\sum_{i=1}^{n} p_i c_{Y(j)_i} = 0 \\
&\iff \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \sum_{k=1}^{p} \alpha_k^* \frac{c_{X_{k(j)i}}}{m} - \sum_{(j)=1}^{m}\sum_{i=1}^{n} p_i \sum_{k=1}^{p} \beta_k^* \frac{c_{X_{k(j)(n-i+1)}}}{m} + \gamma^* = \sum_{(j)=1}^{m}\sum_{i=1}^{n} p_i \frac{c_{Y(j)_i}}{m} \\
&\iff \sum_{k=1}^{p} \left(\alpha_k^* \overline{X_k} - \beta_k^* \overline{X_k} \right) + \gamma^* = \overline{Y}
\end{aligned}$$

From *Property 3.2*, it follows that $\overline{\widehat{Y}} = \sum_{k=1}^{p} (\alpha_k^* - \beta_k^*) \overline{X_k} + \gamma^*$, so $\overline{\widehat{Y}} = \overline{Y}$. □

Property 3.4 *For each unit j, the quantile function for the distribution $\widehat{Y}(j)$ predicted by the DSD Model, can be rewritten as follows:*

$$\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y} = \sum_{k=1}^{p} \alpha_k^* \left(\Psi^{-1}_{X_k(j)}(t) - \overline{X_k} \right) + \beta_k^* \left(-\Psi^{-1}_{X_k(j)}(1-t) + \overline{X_k} \right).$$

Proof: In *Property 3.3*, we proved that

$$\overline{Y} = \sum_{k=1}^{p} (\alpha_k^* - \beta_k^*) \overline{X_k} + \gamma^* \iff \gamma^* = \overline{Y} - \sum_{k=1}^{p} (\alpha_k^* - \beta_k^*) \overline{X_k}.$$

For the optimal solution B^*, for each unit j, the quantile function predicted by the linear regression model *DSD*, in *Definition 3.3*, is given by

$$\Psi_{\widehat{Y}(j)}(t) = \sum_{k=1}^{p} \alpha_k^* \Psi^{-1}_{X_k(j)}(t) - \beta_k^* \Psi^{-1}_{X_k(j)}(1-t) + \gamma^*$$

which may be rewritten as

$$\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y} = \sum_{k=1}^{p} \alpha_k^* \left(\Psi^{-1}_{X_k(j)}(t) - \overline{X_k} \right) + \beta_k^* \left(-\Psi^{-1}_{X_k(j)}(1-t) + \overline{X_k} \right). \quad \square$$

Property 3.5 *For the observed and predicted distributions $Y(j)$ and $\widehat{Y}(j)$, with $j \in \{1,\ldots,m\}$, of the variable Y, we have*

$$\sum_{j=1}^{m} \int_0^1 \left(\Psi_{Y(j)}^{-1}(t) - \Psi_{\widehat{Y}(j)}^{-1}(t)\right)\left(\Psi_{\widehat{Y}(j)}^{-1}(t) - \overline{Y}\right) dt = 0.$$

Proof: The proof is given in *Appendix B*.

3.4 Goodness-of-fit measure

To complete the investigation of the linear regression model for histogram-valued variables, a goodness-of-fit measure remains to be deduced. We define this measure in a similar way as in the classical model for real data.

Proposition 3.2 *The sum of the square of the Mallows distance between each observed distribution j, $j \in \{1,\ldots,m\}$, of the histogram-valued variable Y, and the mean of the histogram-valued variable Y, $\overline{Y}$, can be decomposed as follows:*

$$\sum_{j=1}^{m} D_M^2\left(\Psi_{Y(j)}^{-1}(t), \overline{Y}\right) = \sum_{j=1}^{m} D_M^2\left(\Psi_{Y(j)}^{-1}(t), \Psi_{\widehat{Y}(j)}^{-1}(t)\right) + \sum_{j=1}^{m} D_M^2\left(\Psi_{\widehat{Y}(j)}^{-1}(t), \overline{Y}\right)$$

Proof: Consider each observation j of the histogram-valued variable Y, represented by its quantile function $\Psi_{Y(j)}^{-1}(t)$, and the mean this histogram-valued variable, $\overline{Y}$. We have,

$$\begin{aligned}
&\sum_{j=1}^{m} D_M^2\left(\Psi_{Y(j)}^{-1}(t), \overline{Y}\right) = \sum_{j=1}^{m} \int_0^1 \left(\Psi_{Y(j)}^{-1}(t) - \overline{Y}\right)^2 dt = \\
&= \sum_{j=1}^{m} \int_0^1 \left(\Psi_{Y(j)}^{-1}(t) - \Psi_{\widehat{Y}(j)}^{-1}(t) + \Psi_{\widehat{Y}(j)}^{-1}(t) - \overline{Y}\right)^2 dt = \\
&= \sum_{j=1}^{m} \int_0^1 \left(\Psi_{Y(j)}^{-1}(t) - \Psi_{\widehat{Y}(j)}^{-1}(t)\right)^2 dt + \sum_{j=1}^{m} \int_0^1 \left(\Psi_{\widehat{Y}(j)}^{-1}(t) - \overline{Y}\right)^2 dt + \\
&\quad + 2\sum_{j=1}^{m} \int_0^1 \left(\Psi_{Y(j)}^{-1}(t) - \Psi_{\widehat{Y}(j)}^{-1}(t)\right)\left(\Psi_{\widehat{Y}(j)}^{-1}(t) - \overline{Y}\right) dt
\end{aligned}$$

From *Property 3.5* we have,

$$\sum_{j=1}^{m} \int_0^1 \left(\Psi_{Y(j)}^{-1}(t) - \Psi_{\widehat{Y}(j)}^{-1}(t)\right)\left(\Psi_{\widehat{Y}(j)}^{-1}(t) - \overline{Y}\right) dt = 0.$$

So, we can write

$$\sum_{j=1}^{m} D_M^2\left(\Psi_{Y(j)}^{-1}(t), \overline{Y}\right) =$$

$$= \sum_{j=1}^{m} \int_0^1 \left(\Psi_{Y(j)}^{-1}(t) - \Psi_{\widehat{Y}(j)}^{-1}(t)\right)^2 dt + \sum_{j=1}^{m} \int_0^1 \left(\Psi_{\widehat{Y}(j)}^{-1}(t) - \overline{Y}\right)^2 dt. \qquad \square$$

Therefore, similarly to the classical model, it is possible to define the goodness-of-fit measure of the *DSD Model*.

Definition 3.7 *Consider the observed and predicted distributions of the histogram-valued variable Y and $\widehat{Y}$ represented, respectively, by their quantile functions $\Psi_{Y(j)}(t)$ and $\Psi_{\widehat{Y}(j)}^{-1}(t)$, and the mean of the histogram-valued variable Y, $\overline{Y}$. The goodness-of-fit measure is given by*

$$\Omega = \frac{\sum_{j=1}^{m} D_M^2\left(\Psi_{\widehat{Y}(j)}^{-1}(t), \overline{Y}\right)}{\sum_{j=1}^{m} D_M^2\left(\Psi_{Y(j)}^{-1}(t), \overline{Y}\right)}.$$

In classical linear regression, the coefficient of determination $\mathtt{R}^2$ ranges from 0 to 1. In this case, the goodness-of-fit measure, Ω, also ranges from 0 to 1.

Proposition 3.3 *The goodness-of-fit measure Ω ranges from 0 to 1.*

Proof: Consider the goodness-of-fit measure $\Omega = \dfrac{\sum_{j=1}^{m} D_M^2\left(\Psi_{\widehat{Y}(j)}^{-1}(t), \overline{Y}\right)}{\sum_{j=1}^{m} D_M^2\left(\Psi_{Y(j)}^{-1}(t), \overline{Y}\right)}$. This measure is non-negative.

So, $\Omega \geq 0$.

From *Proposition 3.2*, we have

$$\sum_{j=1}^{m} D_M^2\left(\Psi_{Y(j)}^{-1}(t), \overline{Y}\right) =$$

$$= \sum_{j=1}^{m} \int_0^1 \left(\Psi_{Y(j)}^{-1}(t) - \Psi_{\widehat{Y}(j)}^{-1}(t)\right)^2 dt + \sum_{j=1}^{m} \int_0^1 \left(\Psi_{\widehat{Y}(j)}^{-1}(t) - \overline{Y}\right)^2 dt \iff$$

$$\iff 1 = \frac{\sum_{j=1}^{m} \int_0^1 \left(\Psi_{Y(j)}^{-1}(t) - \Psi_{\widehat{Y}(j)}^{-1}(t)\right)^2 dt}{\sum_{j=1}^{m} D_M^2\left(\Psi_{Y(j)}^{-1}(t), \overline{Y}\right)} + \frac{\sum_{j=1}^{m} \int_0^1 \left(\Psi_{\widehat{Y}(j)}^{-1}(t) - \overline{Y}\right)^2 dt}{\sum_{j=1}^{m} D_M^2\left(\Psi_{Y(j)}^{-1}(t), \overline{Y}\right)}$$

$$\iff \Omega = 1 - \frac{\sum_{j=1}^{m} \int_0^1 \left(\Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)\right)^2 dt}{\sum_{j=1}^{m} D^2_M\left(\Psi^{-1}_{Y(j)}(t), \overline{Y}\right)}$$

Since the term $\dfrac{\sum_{j=1}^{m} \int_0^1 \left(\Psi^{-1}_{Y(j)}(t) - \Psi^{-1}_{\widehat{Y}(j)}(t)\right)^2 dt}{\sum_{j=1}^{m} D^2_M\left(\Psi^{-1}_{Y(j)}(t), \overline{Y}\right)}$ is non-negative, the value of Ω is always less than or equal to 1. So, we have that $0 \leq \Omega \leq 1$.

Let us now analyze the extreme situations.

Suppose $\Omega = 0$. In this case,

$$\sum_{j=1}^{m} D^2_M\left(\Psi^{-1}_{\widehat{Y}(j)}(t), \overline{Y}\right) = 0 \iff \sum_{j=1}^{m} \int_0^1 \left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y}\right)^2 dt = 0.$$

So, for all $j \in \{1, \ldots, m\}$, we have $\Psi^{-1}_{\widehat{Y}(j)}(t) - \overline{Y} = 0 \iff \Psi^{-1}_{\widehat{Y}(j)}(t) = \overline{Y}$. In this case the predicted function for all observations j is a constant function.

Suppose now that $\Omega = 1$. In this case,

$$\sum_{j=1}^{m} D^2_M\left(\Psi^{-1}_{\widehat{Y}(j)}(t), \overline{Y}\right) = \sum_{j=1}^{m} D^2_M\left(\Psi^{-1}_{Y(j)}(t), \overline{Y}\right).$$

From the decomposition obtained in *Proposition 3.2* we have,

$$\sum_{j=1}^{m} D^2_M\left(\Psi^{-1}_{Y(j)}(t), \overline{Y}\right) = \sum_{j=1}^{m} D^2_M\left(\Psi^{-1}_{\widehat{Y}(j)}(t), \overline{Y}\right) + \sum_{j=1}^{m} D^2_M\left(\Psi^{-1}_{\widehat{Y}(j)}(t), \Psi^{-1}_{Y(j)}(t)\right)$$
$$\iff \sum_{j=1}^{m} D^2_M\left(\Psi^{-1}_{\widehat{Y}(j)}(t), \Psi^{-1}_{Y(j)}(t)\right) = 0.$$

So, for all $j \in \{1, \ldots, m\}$,

$$D^2_M\left(\Psi^{-1}_{\widehat{Y}(j)}(t), \Psi^{-1}_{Y(j)}(t)\right) = 0 \iff \int_0^1 \left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \Psi^{-1}_{Y(j)}(t)\right)^2 dt = 0 \Longrightarrow \Psi^{-1}_{\widehat{Y}(j)}(t) = \Psi^{-1}_{Y(j)}(t).$$

In this case, for each observation j, the predicted and observed quantile functions are coincident.

In conclusion $0 \leq \Omega \leq 1$. If $\Omega = 0$ there is no linear relationship between the histogram-valued variable Y and the histogram-valued variables X_k. If $\Omega = 1$, the linear relation is perfect, so the relationship between the histogram-valued variable Y and histogram-valued variables X_k, with $k \in \{1, \ldots, p\}$, is exactly the relation defined by the linear regression model. □

4 Experiments

To illustrate and analyze the *DSD Model* we performed a simulation study and applied the method to real datasets.

4.1 Simulation study

To analyze the behavior of the parameter estimation and the performance of the *DSD Model* in different situations, we performed a simulation study. The first step was to generate the observations of the histogram-valued variables X_k, $k = \{1, \ldots, p\}$ and Y, where Y is the variable to be modelized from X_k by the linear relationship. Next, the parameters were estimated by the *DSD Model* and goodness-of-fit measures computed, considering symbolic simulated data tables covering different situations. From these results it was possible to analyze the behavior of the model and draw some meaningful conclusions.

4.1.1 Building symbolic simulated data tables

The observations of the explicative and response histogram-valued variables X_k and Y were generated in different ways.

- The observations of each histogram-valued variable X_k are created.
  According to the concept of symbolic variables, to obtain the m observations associated to a histogram-valued variable X_k, we started by simulating 5000 real values corresponding to each unit. These values are then organized in histograms, that represent the empirical distribution for each unit. It was considered, without loss of generality, that in all observations, the subintervals of each histogram have the same weight (equiprobable) with frequency 0.10. This option is not restrictive, and is also supported by the work of Colombo [9]. If we had not considered equiprobable histograms with the same weight in all observations, we would have obtained a large number of different weights and consequently the subintervals would have very low frequencies. It is possible that histograms are not equiprobable, however, the weight in each subinterval has to be the same in all observations (see *Subsection 2.1*). Furthermore diversity of weights would lead to rounding errors that increase the difficulty to work with histograms.

- The observations of the histogram-valued variable Y are created.
  The histograms that are the observations of the histogram-valued variable Y are obtained in three steps. First, we consider the perfect linear regression, without error, given by

$$\Psi^{-1}_{Y^*(j)}(t) = \gamma + \sum_{k=1}^{p} \alpha_k \Psi^{-1}_{X_k(j)}(t) - \sum_{k=1}^{p} \beta_k \Psi^{-1}_{X_k(j)}(1-t),$$

  for particular values of the parameters. The histogram-valued variables X_k and Y^* are in a perfect linear relationship, this is however not what is intended to simulate a symbolic data table. Then, we disturb the perfect linear relationship by introducing an error function in the model $\Psi^{-1}_{Y(j)}(t) =$

$\Psi^{-1}_{Y^*(j)}(t) + \varepsilon_j(t)$. The error function is a piece-wise linear function (but not necessarily a quantile function) defined by:

$$\varepsilon_j(t) = \begin{cases} a_{(j)_1} + \left(2\frac{t}{w_1} - 1\right) b_{(j)_1} & if \quad 0 \leq t < w_1 \\ a_{(j)_1} + b_{(j)_1} + b_{(j)_2} + \left(2\frac{t-w_1}{w_2-w_1} - 1\right) b_{(j)_2} & if \quad w_1 \leq t < w_2 \\ \vdots & \\ a_{(j)_1} + b_{(j)_1} + \sum_{i=2}^{n-1} 2b_{(j)_i} + b_{(j)_n} \left(2\frac{t-w_{(n-1)}}{1-w_{(n-1)}} - 1\right) b_{(j)_n} & if \quad w_{n-1} \leq t \leq 1 \end{cases} \tag{17}$$

Each quantile function $\Psi^{-1}_{Y^*(j)}(t)$ is randomly disturbed by the error function for different values of $a_{(j)_1}$ and $b_{(j)_i}$, $i \in \{1, \ldots, n\}$. These values might have a high or low variation depending on whether we want the linear regression between the variables to be better or worse. The selection of these values takes into account the "magnitude" of the values considered in each distribution $\Psi^{-1}_{Y^*(j)}(t)$. The values of $b_{(j)_i}$, cannot be lower than the minimum value of the half range $-r_{Y(j)_i}$, else for this unit j and subinterval i, the half range $r_{Y^*(j)_i}$ would be negative.

To perform the simulation study, symbolic data tables that illustrate different situations were created. For each situation considered, 1000 data tables were generated. In this study a full factorial design was employed, with the following factors:

- Number of explicative histogram-valued variables: $p = 1$ and $p = 3$.
- Parameters of the *DSD Model*.
  - For $p = 1$:
    - **i)** $\alpha = 2$; $\beta = 1$; $\gamma = -1$; (α and β are close)
    - **ii)** $\alpha = 2$; $\beta = 8$; $\gamma = 3$; (α is lower than β)
    - **iii)** $\alpha = 8$; $\beta = 0$; $\gamma = 4$; (α is higher than β)
  - For $p = 3 : \alpha_1 = 2$; $\beta_1 = 1$; $\alpha_2 = 0.5$; $\beta_2 = 3$; $\alpha_3 = 4$; $\beta_3 = 2$; $\gamma = -1$;
- Distribution of the microdata that allow generating the histograms corresponding to each observation of the variables $X_k, \quad k = \{1, \ldots, p\}$:

  i) Uniform distribution

  $(X_k(j) \sim \mathcal{U}(\delta_1(j), \delta_2(j))$ where for each $j \in \{1, \ldots, m\}$, $\delta_1(j) \sim \mathcal{U}(-2, 0)$ and $\delta_2(j) \sim \mathcal{U}(0, 2))$;

  ii) Normal distribution

  $(X_k(j) \sim \mathcal{N}(\mu(j), \sigma^2(j))$ where for each $j \in \{1, \ldots, m\}$, $\mu(j) \sim \mathcal{U}(0, 1)$ and $\sigma^2(j) \sim \mathcal{U}(0, 2))$;

  iii) Log-Normal distribution

  $(X_k(j) \sim ln\mathcal{N}(\mu(j), \sigma^2(j))$ where for each $j \in \{1, \ldots, m\}$, $\mu(j) \sim \mathcal{U}(-0.5, 0.5)$ and $\sigma^2(j) \sim \mathcal{U}(0.5, 1))$;

iv) Mixture of distributions, randomly selected from {Uniform:$X_k(j) \sim \mathcal{U}(1,3)$; Normal: $X_k(j) \sim \mathcal{N}(1,1)$; Chi-square: $X_k(j) \sim \mathcal{X}^2(1)$; Log-normal: $X_k(j) \sim ln\mathcal{N}(0,0.5)$; -Log-normal: $X_k(j) \sim -ln\mathcal{N}(0,0.5)$}

- Level of the linearity of the model:

  i) High linearity - In the error $\varepsilon_j(t)$, the values of $a_{(j)_1}$ and $b_{(j)_i}$ are randomly generated in $\mathcal{U}_{c1} = \mathcal{U}(-\frac{3}{8} * \ddot{C}, \frac{3}{8} * \ddot{C})$[1] and $\mathcal{U}_{r1} = \mathcal{U}(-\frac{1}{8} * min(r_{Y^*(j)_i}), \frac{1}{8} * min(r_{Y^*(j)_i}))$, respectively;

  ii) Moderate linearity - In the error $\varepsilon_j(t)$, the values of $a_{(j)_1}$ and $b_{(j)_i}$ are randomly generated in $\mathcal{U}_{c2} = \mathcal{U}(-\frac{3}{2} * \ddot{C}, \frac{3}{2} * \ddot{C})$ and $\mathcal{U}_{r2} = \mathcal{U}(-\frac{1}{2} * min(r_{Y^*(j)_i}), \frac{1}{2} * min(r_{Y^*(j)_i}))$, respectively;

  iii) Low linearity - In the error $\varepsilon_j(t)$, the values of $a_{(j)_1}$ and $b_{(j)_i}$ are randomly generated in $\mathcal{U}_{c3} = \mathcal{U}(-3 * \ddot{C}, 3 * \ddot{C})$ and $\mathcal{U}_{r3} = \mathcal{U}(-min(r_{Y^*(j)_i}), min(r_{Y^*(j)_i}))$, respectively;

- Sample size: m=10; 30; 100; 250.

It is important to underline that in this simulation study, it was only possible to control the type of distributions in observations of the explicative histogram-valued variables. This simulation does not allow selecting the distributions in the observations of the response variable. These distributions depend of the distribution of the variables Y^* (that in some situations are known, as we will see later) and the disturbance applied to the histograms $Y^*(j)$.

4.1.2 Description of the simulation study

The simulated symbolic data tables include the observations of the histogram-valued variables X_k and Y, according to the previous description and factors. For these tables, we computed the estimated parameters for the *DSD Model* and the goodness-of-fit measures. As we considered 1000 replications for each situation, the values presented are the means of the obtained values and the respective standard deviation values (represented by s).

The goodness-of-fit measures considered in this study are :

- $\overline{\Omega}$, where Ω is the measure deduced from the *DSD Model* (see *Subsection 3.4*);
- Root-mean-square error ($RMSE_M$), a measure defined using the Mallows distance (also used in the *DSD Model*), proposed by Irpino and Verde [13]; it is defined by

$$RMSE_M = \sqrt{\frac{\sum_{j=1}^{m} \int_0^1 \left(\Psi^{-1}_{\widehat{Y}(j)}(t) - \Psi^{-1}_{Y(j)}(t)\right)^2 dt}{m}}$$

[1] $\ddot{C} = \frac{1}{n} \sum_{i=1}^{n} \left| \sum_{j=1}^{m} \frac{1}{m} c_{Y(j)_i} \right|$

- Adaptations of the lower ($RMSE_L$) and the upper bound ($RMSE_U$) root-mean-square that Neto and Carvalho [17], [18], use to study the performance of the linear regression models defined for interval-valued variables; for histogram-valued variables, the $RMSE_L$ and the $RMSE_U$ are given by:

$$RMSE_L = \frac{1}{m}\sum_{j=1}^{m}\sqrt{\sum_{i=1}^{n}(\underline{I}(j)_i - \widehat{\underline{I}}(j)_i)^2 p_i} \qquad RMSE_U = \frac{1}{m}\sum_{j=1}^{m}\sqrt{\sum_{i=1}^{n}(\overline{I}(j)_i - \widehat{\overline{I}}(j)_i)^2 p_i}$$

with $\left[\underline{I}(j)_i, \overline{I}(j)_i\right[$ and $\left[\widehat{\underline{I}}(j)_i, \widehat{\overline{I}}(j)_i\right[$ the subintervals $i \in \{1, \ldots, n\}$ of the observed and predicted histograms, for each unit j.

In *Appendix C* four tables are presented, each of which containing the results obtained with $p = 1$ and all distributions used for defining the histogram values of X_j, i.e., all observations with Uniform distribution (*Table 6*), Normal distribution (*Table 7*), Log-Normal distribution (*Table 8*) and the observations of $X(j)$ for a mixture of distributions (*Table 9*). In the last two tables, similar results are presented for the cases where $p = 3$ (*Table 10*, *Table 11*).

4.1.3 Results and conclusions

The main goals of this study are to analyze the behavior of the parameters' estimation and the performance of the *DSD Model*. The results obtained for the model with one or three explicative variables are similar, and as such in this subsection we will only be analyzing with detail the results obtained when $p = 1$. The results obtained when $p = 3$ may be found in *Table 10* and *Table 11* of *Appendix C*. For $p = 1$ it is also our goal to analyze how the symmetry/asymmetry of the distributions in observations of the explicative histogram-valued variable affect the symmetry/asymmetry of the distributions in observations of the predictive variable.

Concerning the analysis of the parameters' estimation.
For the simple case with one explicative histogram-valued variable X, we considered the mean of the values obtained for $\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}$ and the mean square error (MSE) [17]. In this case, as we replicated the same situation 1000 times, $MSE = \frac{1}{1000}\sum_{i=1}^{1000}(\theta - \widehat{\theta}_i)^2$ (with θ corresponding to each parameter of the model). Comparing the first four tables in *Appendix C*, we can see that the behavior of the parameters' estimation is independent of the distribution used to generate the microdata of the explicative variables. Futhermore, the estimated parameters are almost always close to the initial parameter values irrespecive of the level of the linearity. This behavior is expected when the level of linearity is higher but when the level of linearity is moderate or low it would not be surprising if the estimated parameters were more distant from the original ones, since it seems natural that other models exist that better adjust the symbolic data. This is essentially observed in *Tables 6 and 7*, when the initial parameters α and β are not close and when the number of observations is lower. According to this, the analysis of the behavior of the

MSE and the mean of the estimated parameters is essentially applicable in situations where the level of linearity is high. For these cases we observe that the values of the MSE decrease and tend to zero as the number of observations increases and the mean of the estimated parameters becomes very close to the respective parameters of the model. These results confirm the empirical consistency of the estimators. In the boxplots presented in *Figures 6, 7* and *8* we may observe that, considering the different types of distributions used to generate the histogram values of X, the boxes reduced their ranges around the true values of the respective parameters as the number m of observations increases. The figures illustrate only the situation when $\alpha = 2; \beta = 1; \gamma = -1$, but the behavior for the other values is similar. It can also be observed that the range of the variation of the estimated parameters relatively to its original value is influenced by the distribution of the observations.

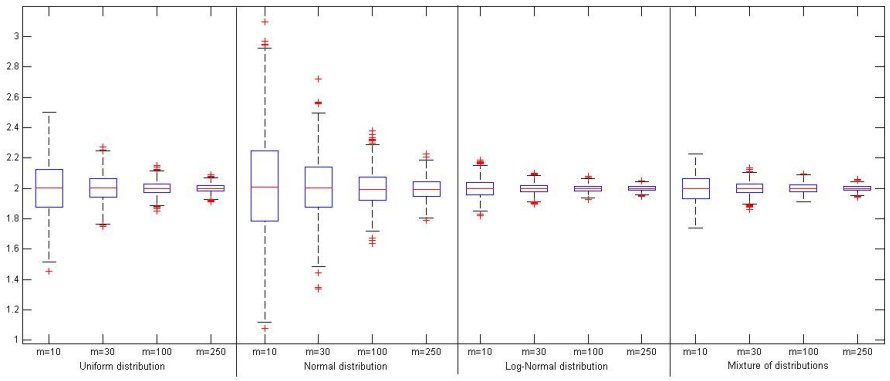

Figure 6: Boxplot for the estimated parameter $\widehat{\alpha}$ for a high level of linearity.

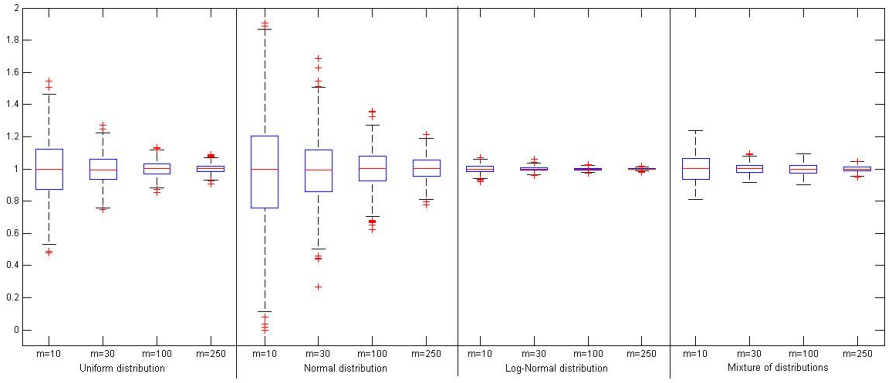

Figure 7: Boxplot for the estimated parameter $\widehat{\beta}$ for a high level of linearity.

Concerning the study of the goodness-of-fit measures.
The values obtained for Ω show that this value provides a good evaluation for the level of linearity. The models slightly disturbed presented values of Ω close to one. On the other hand, when the error function applied to the model presented a high level of variability the values of Ω are closer to zero. Furthermore,

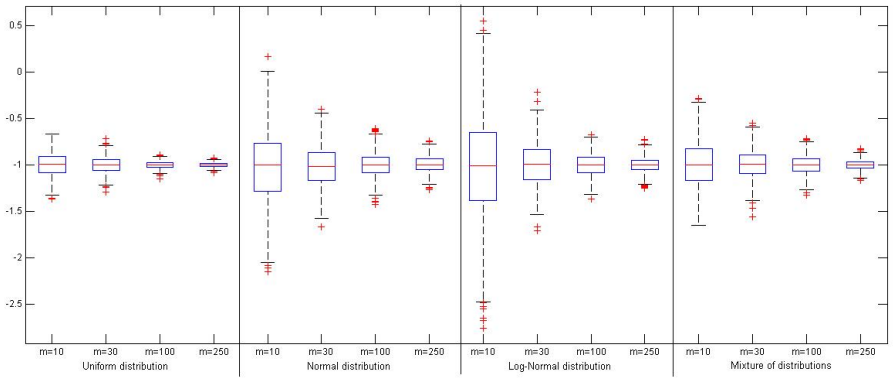

Figure 8: Boxplot for the estimated parameter $\widehat{\gamma}$ for a high level of linearity.

the means of the values Ω are consistent with the respective values of the measures $RMSE_M$; $RMSE_L$ and $RMSE_U$. In general, as expected, the highest values of the Ω correspond to the lowest values of $RMSE_M$. In all tables of *Appendix C* we can also verify that in almost all situations, the values of the goodness of fit measure decrease in the same proportion as the levels of linearity. The level of linearity and the mean values associated to the goodness-of-fit measures $RMSE_M$; $RMSE_L$ and $RMSE_U$ increase approximately four times when we pass from high to moderate linearity and approximately two times when we pass from moderate to low. This increase is an exact reflection of the range of variability tested in this study for the error function (four times from the high to moderate linearity and two times from moderate to low).

Tables 2 and *3* illustrate the results that were obtained in an additional study for the original model with $\alpha = 2; \beta = 1; \gamma = -1$ and only for samples with 10 and 100 observations. Other situations were tested and the results were similar. The goal of the study was to analyze the level of sensitivity of the measure Ω to different kinds of error functions, that in some cases affect more the half range of the subintervals of the histograms and in others the centers. To analyze this behavior, the values of Ω were determined, considering different error functions that use three levels of variability for the values of $a_{(j)_1} : \mathcal{U}_{c1}, \mathcal{U}_{c2}, \mathcal{U}_{c3}$ as defined in *Subsection 4.1.1* and, for each one, three levels of variability for $b_{(j)_i} : \mathcal{U}_{r1}, \mathcal{U}_{r2}, \mathcal{U}_{r3}$ as defined in *Subsection 4.1.1*.

		$\overline{\Omega}_{\mathcal{U}}\ (s)$			$\overline{\Omega}_{\mathcal{N}}\ (s)$		
	m	$b_{(j)_i} \sim \mathcal{U}_{r1}$	$b_{(j)_i} \sim \mathcal{U}_{r2}$	$b_{(j)_i} \sim \mathcal{U}_{r3}$	$b_{(j)_i} \sim \mathcal{U}_{r1}$	$b_{(j)_i} \sim \mathcal{U}_{r2}$	$b_{(j)_i} \sim \mathcal{U}_{r3}$
$a_{(j)_1} \sim \mathcal{U}_{c1}$	10	0.9741 (0.0089)	0.9455 (0.0216)	0.8643 (0.0535)	0.9792 (0.0079)	0.9145 (0.0344)	0.7587 (0.0835)
	100	0.9648 (0.0032)	0.9322 (0.0076)	0.8403 (0.0191)	0.9762 (0.0025)	0.8982 (0.0122)	0.7160 (0.0293)
$a_{(j)_1} \sim \mathcal{U}_{c2}$	10	0.7323 (0.0727)	0.7163 (0.0786)	0.6690 (0.0906)	0.7980 (0.0583)	0.7567 (0.0691)	0.6555 (0.0997)
	100	0.6476 (0.0222)	0.6332 (0.0238)	0.5905 (0.0288)	0.7701 (0.0165)	0.7217 (0.0232)	0.6008 (0.0326)
$a_{(j)_1} \sim \mathcal{U}_{c3}$	10	0.4422 (0.1098)	0.4320 (0.1090)	0.4211 (0.1120)	0.5192 (0.0931)	0.5017 (0.0944)	0.4587 (0.1013)
	100	0.3195 (0.0251)	0.3156 (0.0265)	0.3054 (0.0268)	0.4627 (0.0243)	0.4436 (0.0260)	0.3970 (0.0306)

Table 2: Mean values of Ω considering different levels of linearity, when the distributions generating observations of X are Uniform $\left(\overline{\Omega}_{\mathcal{U}}\right)$ and Normal $\left(\overline{\Omega}_{\mathcal{N}}\right)$.

		$\overline{\Omega}_{Ln\mathcal{N}}\ (s)$			$\overline{\Omega}_{M}\ (s)$		
	m	${}^{b_{(j)_i}}\sim \mathcal{U}_{r1}$	${}^{b_{(j)_i}}\sim \mathcal{U}_{r2}$	${}^{b_{(j)_i}}\sim \mathcal{U}_{r3}$	${}^{b_{(j)_i}}\sim \mathcal{U}_{r1}$	${}^{b_{(j)_i}}\sim \mathcal{U}_{r2}$	${}^{b_{(j)_i}}\sim \mathcal{U}_{r3}$
${}^{a_{(j)_1}}\sim \mathcal{U}_{c1}$	10	0.9843 (0.0054)	0.8848 (0.0389)	0.6699 (0.0994)	0.9780 (0.0078)	0.9203 (0.0275)	0.7789 (0.0721)
	100	0.9822 (0.0019)	0.8786 (0.0146)	0.6571 (0.0393)	0.9719 (0.0026)	0.9042 (0.0095)	0.7403 (0.0220)
${}^{a_{(j)_1}}\sim \mathcal{U}_{c2}$	10	0.8769 (0.0344)	0.8032 (0.0587)	0.6130 (0.1092)	0.7765 (0.0569)	0.7453 (0.0706)	0.6568 (0.0954)
	100	0.8556 (0.0112)	0.7765 (0.0229)	0.5982 (0.0420)	0.7225 (0.0182)	0.6838 (0.0223)	0.5884 (0.0287)
${}^{a_{(j)_1}}\sim \mathcal{U}_{c3}$	10	0.6542 (0.0762)	0.6114 (0.0923)	0.5067 (0.1208)	0.4884 (0.0948)	0.4791 (0.0956)	0.4422 (0.1024)
	100	0.6075 (0.0224)	0.5654 (0.0293)	0.4638 (0.0418)	0.3979 (0.0228)	0.3855 (0.0226)	0.3526 (0.0260)

Table 3: Mean values of Ω considering different levels of linearity when the distributions generating observations of X are Log-Normal $\left(\overline{\Omega}_{Ln\mathcal{N}}\right)$ and a mixture of distributions $\left(\overline{\Omega}_{M}\right)$.

Based on these results, we can say that, except when the observations of the explicative variables follow a Log-Normal distribution, the linearity between histogram-valued variables is more affected by disturbances in the center of the subintervals than in the half range. This behavior is not surprising because the distance associated to this model is the the Mallows distance and as we observe in its definition the contribution to the centers of the subintervals is three times more then that of the half-ranges (see *Definition 3.1* and *Property 3.1*). On the other hand, when all observations of the explicative histogram-valued variable have asymmetric distributions, the influence of the disturbance in the center and half-range may be similar. This different behavior may be related to the kind of distribution (symmetric/asymmetric) predicted for the observations of the histogram-valued variable $\widehat{Y}(j)$, as we will see next.

Concerning symmetry/assymetry of $\widehat{Y}(j)$.
In this simulation study it was possible to analyze the symmetry/assymetry of the predicted distributions obtained by the *DSD Model*, taking into consideration the symmetry/asymmetry of the distributions in the observations of the histogram-valued variables X and the values of the parameters α and β. When the observation of the histogram-valued variable X has a symmetric distribution, represented by $\Psi^{-1}_{X(j)}(t)$, the respective symmetric distribution $-\Psi^{-1}_{X(j)}(1-t)$ is also symmetric, but when the distribution $\Psi^{-1}_{X(j)}(t)$ is asymmetric positive (negative) (Log-Normal, for example), the respective symmetric distribution $-\Psi^{-1}_{X(j)}(1-t)$ is asymmetric negative (positive). In the *DSD Model*, the predicted distributions are obtained from $\Psi^{-1}_{\widehat{Y}(j)}(t) = \gamma + \alpha\Psi^{-1}_{X(j)}(t) - \beta\Psi^{-1}_{X(j)}(1-t)$. Therefore if the distribution $\Psi^{-1}_{X(j)}(t)$ is symmetric the distribution of $\Psi^{-1}_{\widehat{Y}(j)}(t)$ also tends to be symmetrical, if the distributions $\Psi^{-1}_{X(j)}(t)$ is asymmetric the distribution of $\Psi^{-1}_{\widehat{Y}(j)}(t)$ tends to be symmetrical when the values of α and β are close, asymmetrical negative (resp. positive) when the value of α is lower (resp. higher) than the value of β. These conclusions are illustrated in *Figure 9* considering all predicted distributions in the simulation study.

In conclusion, when the distributions of observations $X(j)$ are symmetric; asymmetric positive or asymmetric negative, it is possible to forecast whether the distributions of $\widehat{Y}(j)$ will be symmetric or asymmetric.

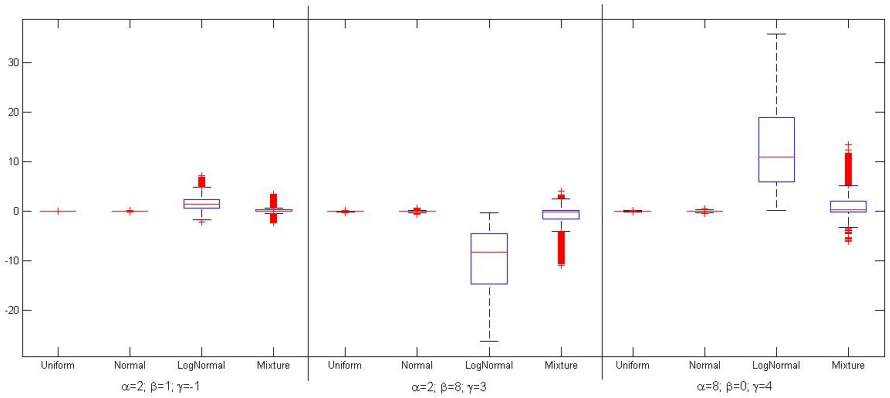

Figure 9: Boxplots of the difference between the mean and median of the estimated distributions in all situations with $n = 10$ considered in the simulation study, for $p = 1$.

4.2 Applied examples

4.2.1 The relation between the hematocrit values and hemoglobin values

This first example was presented in [3] to illustrate their linear regression model for histogram-valued variables. In this case, we have the symbolic data in *Table 4*, where 10 units are described by two symbolic variables, the hematocrit and the hemoglobin.

Obs.	Hematocrit (Y)	Hemoglobin (X)
1	$\{[33.29; 37.52[\ , 0.6; [37.52; 39.61]\ , 0.4\}$	$\{[11.54; 12.19[\ , 0.4; [12.19; 12.8]\ , 0.6\}$
2	$\{[36.69; 39.11[\ , 0.3; [39.11; 45.12]\ , 0.7\}$	$\{[12.07; 13.32[\ , 0.5; [13.32; 14.17]\ , 0.5\}$
3	$\{[36.69; 42.64[\ , 0.5; [42.64; 48.68]\ , 0.5\}$	$\{[12.38; 14.2[\ , 0.3; [14.2; 16.16]\ , 0.7\}$
4	$\{[36.38; 40.87[\ , 0.4; [40.87; 47.41]\ , 0.6\}$	$\{[12.38; 14.26[\ , 0.5; [14.26; 15.29]\ , 0.5\}$
5	$\{[39.19; 50.86]\ , 1\}$	$\{[13.58; 14.28[\ , 0.3; [14.28; 16.24]\ , 0.7\}$
6	$\{[39.7; 44.32[\ , 0.4; [44.32; 47.24]\ , 0.6\}$	$\{[13.81; 14.5[\ , 0.4; [14.5; 15.2]\ , 0.6\}$
7	$\{[41.56; 46.65[\ , 0.6; [46.65; 48.81]\ , 0.4\}$	$\{[14.34; 14.81[\ , 0.5; [14.81; 15.55]\ , 0.5\}$
8	$\{[38.4; 42.93[\ , 0.7; [42.93; 45.22]\ , 0.3\}$	$\{[13.27; 14.0[\ , 0.6; [14.0; 14.6]\ , 0.4\}$
9	$\{[28.83; 35.55[\ , 0.5; [35.55; 41.98]\ , 0.5\}$	$\{[9.92; 11.98[\ , 0.4; [11.98; 13.8]\ , 0.6\}$
10	$\{[44.48; 52.53]\ , 1\}$	$\{[15.37; 15.78[\ , 0.3; [15.78; 16.75]\ , 0.7\}$

Table 4: Example of symbolic data table where the two variables hematocrit and hemoglobin are histogram-valued variables.

We predicted the quantile function representing the distribution taken by the histogram-valued variable Y from the *DSD Model*, and obtained:

$$\Psi^{-1}_{\widehat{Y}(j)}(t) = -1.953 + 3.5598\Psi^{-1}_{X(j)}(t) - 0.4128\Psi^{-1}_{X(j)}(1-t)$$

The value of the goodness-of-fit measure is, for this case, $\Omega = 0.96$.

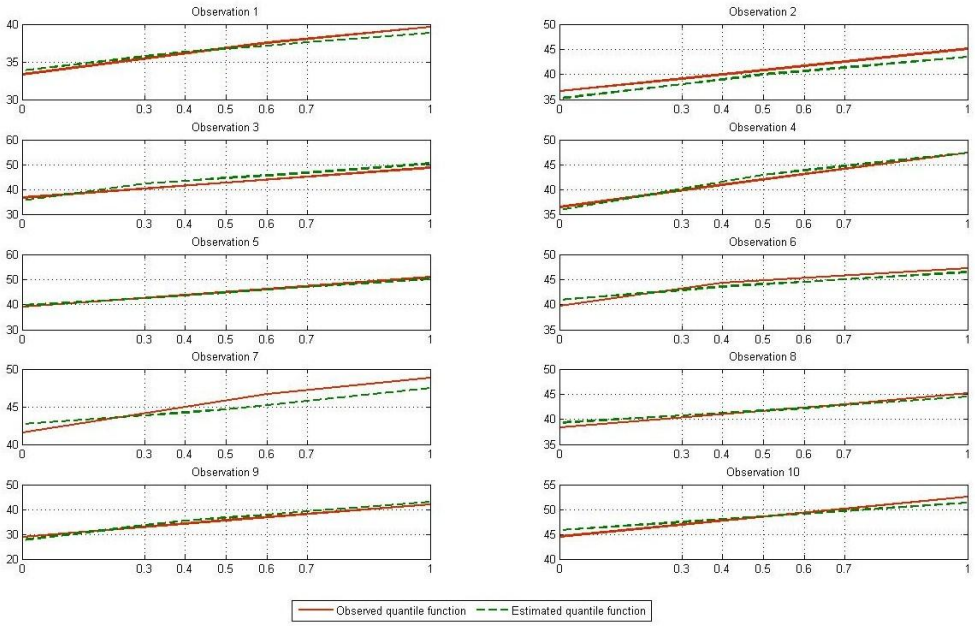

Figure 10: Observed and predicted quantile functions of each observation in *Table 4*.

In *Figure 10* we may compare the quantile functions of the observed and predicted distributions of the histogram-valued variable Y. As it may be observed, the distributions are very similar, in agreement with the value of the coefficient of determination, Ω. The observed and predicted histograms of each observation are presented in *Appendix D*.

When we predict a histogram value we have always associated an error function defined according to *Definition 3.3*. For this example, in *Figure 11* we can observe the error function for observations 1 and 3.

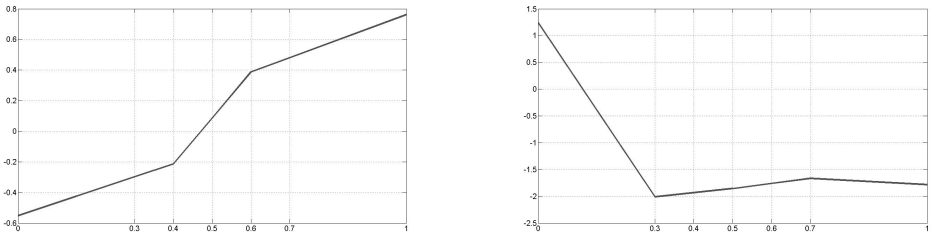

Figure 11: Error function for the observations 1 and 3.

The relationship between the histogram-valued variables in *Table 4* may be visualized in the scatter plot for histograms in *Figure 12*. In this graphic, each of the distributions is represented by a histogram with a different color. These graphics show that a strong linear relation between the histogram-valued variables hematocrit and hemoglobin is observed.

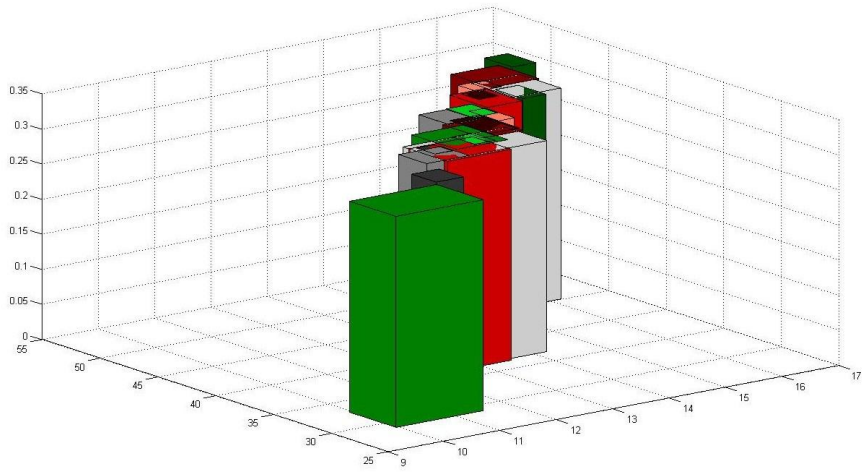

Figure 12: Scatter plot of the data in *Table 4*.

From *Property 3.2* we may conclude that for the set of patients to which the data refers, the symbolic mean of hematocrit increases $\alpha - \beta = 3.1470$ for each unit of increase of the symbolic mean of hemoglobin. As this value is positive we may consider that the relationship between the histogram-valued variables is direct.

For this example, we also predicted the hematocrit distributions using the linear regression models proposed by Billard and Diday [5] and Irpino and Verde [13],[23]. The hematocrit distributions obtained by these methods are presented in *Appendix D*. To compare the performance of the methods, the measures $RMSE_M, RMSE_L, RMSE_U$ (see *Subsection 4.1.3*) were used (see *Table 5*).

Measure	*DSD Model*	*Billard-Diday Model*	*Verde-Irpino Model*
$RMSE_L$	0.8806	1.0288	0.9220
$RMSE_U$	0.8432	1.1064	0.8645
$RMSE_M$	0.8946	1.0507	0.9145

Table 5: Comparison of the performance between the *DSD Model*, the *Billard-Diday Model* and the *Verde-Irpino Model*.

4.2.2 Distributions of Crimes in USA

In this example we consider a real data table (microdata) [20] where we have records related with communities in the USA. The original data combines socio-economic data from the '90 Census and crime data from 1995. For this study we selected the response variable *violent crimes* (total number of violent crimes per 100 000 habitants) and four explicative variables: X_1 (percentage of people aged 25 and over with less than 9th grade education); X_2 (percentage of people aged 16 and over who are employed); X_3 (percentage of population who are divorced); X_4 (percentage of immigrants who immigrated within the

last 10 years). To build the symbolic data table we aggregated the information (contemporary aggregation) for each state. The units (higher units) of this study are the states of USA and their observations for each selected variable are the distributions of the records of the communities of the respective state. To build the initial data table we considered only the states for which the number of records for the variables selected was higher than thirty. Using this criterion, only twenty states were included (AL, CA, CT, FL, GA, IN, MA, MO, NC, NJ, NY, OH, OK, OR, PA, TN, TX, VA, WA, WI). Similarly to the simulation study, we consider, without loss of generality, that in all observations, the subintervals of each histogram have the same weight (equiprobable) with frequency 0.20. Furthermore as the response variable *violent crimes* admits only positive values and the distributions of these values are asymmetric, we will consider as response histogram-valued variable, the variable LVC whose observations are the distributions of the logarithm of the number of violent crimes for each USA state. Considering these conditions, the model that allows to predict the distribution of LVC from the distributions of the explicative variables X_1, X_2, X_3 and X_4, for each USA state j is as follows:

$$\begin{aligned}\Psi^{-1}_{\widehat{LVC}(j)}(t) &= 3.9321 + 0.0009\Psi^{-1}_{X_1(j)}(t) - 0.0123\Psi^{-1}_{X_2(j)}(1-t) + \\ &\quad +0.2073\Psi^{-1}_{X_3(j)}(t) - 0.0353\Psi^{-1}_{X_3(j)}(1-t) + 0.0187\Psi^{-1}_{X_4(j)}(t) \qquad (18)\end{aligned}$$

with $t \in [0,1]$. The goodness-of-fit measure associated to this model is $\Omega = 0.87$.

The values of the parameters estimated for this situation allow to conclude that the variables X_1, X_3 and X_4 have a direct influence in the logarithm of the number of violent crimes and the percentage of employed people have an opposite effect. From *Property 3.2* we may conclude that, for the set of states to which the data refer, when the symbolic mean of the percentage of population divorced increases 1% and the other variables remain constant, the symbolic mean of the LVC increases 0.1720. The percentage of divorced population is the one that influences the most the predicted histogram-valued variable. This interpretation can be extrapolated for the values of the associated parameter of all other explicative variables.

The advantage of studying a linear relationship between data with variability is the possibility to predict the distribution of the values of the response variable instead of only one real value as in a classical study. In this example, the predicted distribution of the logarithm of the number of violent crimes for a given state is more informative about the criminality in that state than only one descriptive measure (e.g., the mean).

Consider one state that was not used to build the model, the state of Arkansas (AR). It is possible to predict the distribution of LVC if the distributions of the explicative variables for this state are known. The histogram predicted by the *DSD Model* (18) for the state Arkansas is

$$H_{LVC}(AR) = \{[4.2250, 5.3158], 0.2; [5.3158, 5.8887], 0.2; [5.8887, 6.4802], 0.2; [6.4802, 7.0509], 0.2; [7.0509, 7.7913], 0.2\}$$

Figure 13 illustrates the estimated and observed quantile function for this state and the values of the measures $RMSE_M, RMSE_L, RMSE_U$ (see *Subsection 4.1.3*). The values of the goodness-of-fit measures prove the closeness between the observed and estimated quantile function that we may see in the figure.

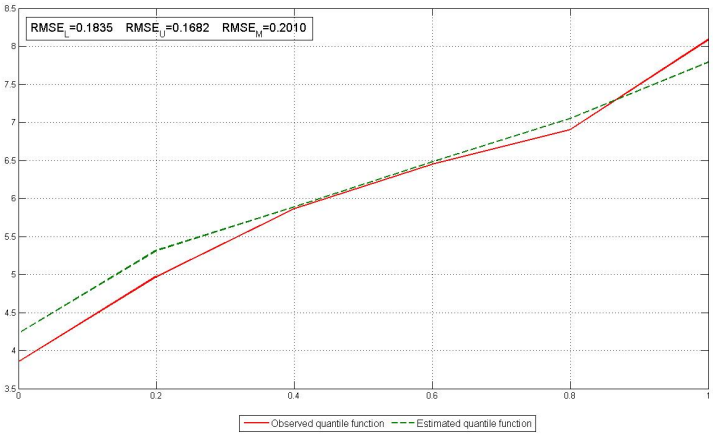

Figure 13: Observed and estimated quantile function of the variable LVC in the state of Arkansas

Analyzing the predicted distribution, we may conclude that in the state of Arkansas the estimated distribution tends to an uniform behavior with the values of LVC to range between 4.23 and 7.79.

The classical alternative to study the logarithm of the number of violent crimes in each USA state would be to reduce the records of all communities of each state, for example to the mean value and make a classical linear regression study. In this case, the variability of the records would be lost and the predicted results would be less informative. Considering the mean of the records associated to each community, the classical model is the following:

$$\widehat{\overline{LVC}}(j) = 6.5817 + 0.0705\overline{X}_1(j) - 0.0503\overline{X}_2(j) + 0.0933\overline{X}_3(j) + 0.0177\overline{X}_4(j) \qquad (19)$$

For this model the value of $\mathrm{R}^2 = 0.75$.

Considering again the state of Arkansas, with the previous model (19) the estimative for $\widehat{\overline{LVC}}(AR)$ is 6.4511. With this approach the information about the behavior of the predicted variable is obviously poorer.

5 Conclusion and perspectives

The *DSD Model* allows predicting the distributions taken by one histogram-valued variable from the distributions taken by explicative histogram-valued variables. Moreover, it is possible to deduce a goodness-of-fit measure from the model. This measure appears to have a good behavior: when we compare the representation of the predicted and observed quantile functions for each unit we have good estimates when the value of the goodness-of-fit measure is close to one whereas the predicted and observed quantile functions are more discrepant when the value of the goodness-of-fit measure is lower. As interval-valued variables are a particular case of histogram-valued variables it is possible to particularize this model for interval-valued variables. An extension of the *DSD Model*, where instead of a real number we use a quantile function as the independent parameter, is under development. This approach will be applied both to histogram-valued variables and interval-valued variables. With this new approach we expect to obtain a more flexible model. Finally, and as a future research perspective, other models and methods in Symbolic Data Analysis based on linear relationships between variables may now be developed using this approach.

Acknowledgments
This work is funded by the ERDF European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP - 01-0124-FEDER-022701.

References

[1] Arroyo, J., Maté, C., 2009. Forecasting Histogram Time Series with K-Nearest Neighbours Methods. International Journal of Forecasting. 25, 192-207.

[2] Arroyo J., 2008. Métodos de Predicción para Series Temporales de Intervalos e Histogramas. Tesis para la obtención del título de Doctor, Universidad Pontificia Comillas, Madrid.

[3] Billard, L., Diday, E., 2006. Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley & Sons, Ltd., Chichester.

[4] Billard, L., Diday, E., 2003. From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. Journal of the American Statistical Association. 98 (462), 470-487.

[5] Billard, L., Diday, E., 2002. Symbolic Regression Analysis, In: Jajuga, K., Sokolowski, A., Bock, H.-H. (Eds.), Classification, Clustering and Data Analysis, Proceedings of the Conference of the International Federation of Classification Societies (IFCS02). Springer, Heidelberg, 281-288.

[6] Billard, L., Diday, E., 2000. Regression Analysis for Interval-Valued Data. In: Kiers, H.A.L., Rasson, J.P., Groenen, P.J.F., Schader M. (Eds.), Data Analysis, Classification, and Related Methods. Proceedings of the Conference of the International Federation of Classification Societies (IFCS00). Springer, Heidelberg, 369-374.

[7] Bock, H.-H., Diday, E. (Eds.), 2000. Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data. Springer-Verlag, Berlin-Heidelberg.

[8] Brito, P., Duarte Silva, A. P., 2011. Modelling Interval Data with Normal and Skew-Normal Distributions. Journal of Applied Statistics. 39 (1), 3-20.

[9] Colombo, A., Jaarsma, R., 1980. A Powerful Numerical Method to Combine Random Variables. IEEE Transactions on Reliability. 29 (2), 126-129.

[10] Diamond, P., 1990. Least squares fitting of compact set-valued data. Journal of Mathematical Analysis and Applications.1 47, 531-544.

[11] Diday, E., 1988. The Symbolic Approach in Clustering and Related Methods of Data Analysis: The Basic Choices. In: Bock, H.-H. (Eds.), Classification and Related Methods of Data Analysis, Proceedings of the Conference of the International Federation of Classification Societies (IFCS87). North Holland, Amsterdam, 673-684.

[12] Diday E. and Noirhomme-Fraiture, M. (Eds.), 2008. Symbolic Data Analysis and the SODAS Software. John Wiley & Sons, Ltd., Chichester.

[13] Irpino, A., Verde, R., 2012. Linear regression for numeric symbolic variables: an ordinary least squares approach based on Wasserstein Distance In web: http://arxiv.org/abs/1202.1436v1 (arXiv:1202.1436v1 [stat.ME] 7Feb2012.)

[14] Irpino, A., Verde, R., 2006. A New Wasserstein Based Distance for the Hierarchical Clustering of Histogram Symbolic Data. In: Batagelj, V., Bock, H.-H, Ferligoj, A.(Eds), Classification and Data Analysis, Proceedings of the Conference of the International Federation of Classification Societies (IFCS06). Springer, Heidelberg, 185-192.

[15] Mallows, C. L., 1972. A note on assymptotic joint normality. The Annals of Mathematical Satistics. 43(2), 508-515.

[16] Neto, E.A.L., Cordeiro, G.M., De Carvalho, F.A.T., 2011. Bivariate symbolic regression models for interval-valued variables. Journal of Statistical Computacion and Simulation. iFirst, 1-18.

[17] Neto, E.A.L., De Carvalho, F.A.T., 2010. Constrained Linear Regression Models for Symbolic Interval-Valued Variables. Computational Statistics & Data Analysis. 54, 333-347.

[18] Neto, E.A.L., De Carvalho, F.A.T., 2008. Centre and Range Method for Fitting a Linear Regression Model to Symbolic Intervalar Data. Computational Statistics & Data Analysis. 52, 1500-1515.

[19] Noirhomme-Fraiture, M. and Brito, P., 2011. Far Beyond the Classical Data Models: Symbolic Data Analysis. Statistical Analysis and Data Mining. 4(2), 157-170.

[20] Redmond M., 2011. UCI Machine Learning Repository [http://www.ics.uci.edu/ mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

[21] Rodriguez O., Diday E., Winsberg S., 2000. Generalization of the Principal Components Analysis to Histogram Data. In: Proceedings of the 4th European Conference on Principles and Practice of Knowledge Discovery in Data Bases (PKDD); Workshop on Symbolic Data Analysis; Lyon (France).

[22] Rodriguez O., Pacheco A., 2004. Applications of Histogram Principal Components Analysis. In: The 15th European Conference on Machine Learning (ECML) and the $8th$ European Conference on Principles and Practice of Knowledge Discovery in Data Bases (PKDD), Pisa (Italy).

[23] Verde, R., Irpino, A., 2010. Ordinary Least Squares for Histogram Data Based on Wasserstein Distance. In: Lechevallier, Y., Saporta, G. (Eds.), Proceedings of COMPSTAT'2010. Physica Verlag, Heidelberg, 581-589.

[24] Verde, R.,Irpino, A., 2008. Comparing Histogram Data Using a Mahalanobis-Wasserstein Distance. In: Brito, P. (Eds.), Proceedings of COMPSTAT'2008. Physica Verlag, Heidelberg, 77-89.

[25] Verde, R.,Irpino, A., 2007. Dynamic Clustering of Histogram Data: Using the Right Metric. In: Brito, P., Bertrand, P., Cucumel, G., De Carvalho, F. (Eds.), Selected Contributions in Data Analysis and Classification. Springer, Heidelberg, 123-134.

[26] Williamson, R., 1989. Probabilistic Arithmetic. Thesis for the obtencion the degree of Doctor, Department of Electrical Engineering, University of Queensland, Australia.

[27] Winston, W., 1994. Operations Research. Applications and Algorithms, third Edition. Duxbury Press. California.

Appendix A: First order partial derivatives of the function SE

$$SE = \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left[\left(c_{Y(j)_i} - \sum_{k=1}^{p} \left(\alpha_k c_{X_k(j)_i} - \beta_k c_{X_k(j)_{n-i+1}} \right) - \gamma \right)^2 + \frac{1}{3} \left(r_{Y(j)_i} - \sum_{k=1}^{p} \left(\alpha_k r_{X_k(j)_i} + \beta_k r_{X_k(j)_{n-i+1}} \right) \right)^2 \right]$$

In these partial derivatives the subintervals of the histograms are defined from the center and half-range of the intervals.

$$\begin{aligned}
\frac{\partial SE(B)}{\partial \alpha_{\mathbf{k}}} &= \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left[2 \left(c_{Y(j)_i} - \gamma - \sum_{k=1}^{p} \left(\alpha_k c_{X_k(j)_i} + \beta_k \left(-c_{X_k(j)_{n-i+1}} \right) \right) \right) \left(-c_{X_{\mathbf{k}}(j)_i} \right) + \frac{2}{3} \left(r_{Y(j)_i} - \sum_{k=1}^{p} \left(\alpha_k r_{X_k(j)_i} + \beta_k r_{X_k(j)_{n-i+1}} \right) \right) \left(-r_{X_{\mathbf{k}}(j)_i} \right) \right] = \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(2 \sum_{k=1}^{p} c_{X_k(j)_i} c_{X_{\mathbf{k}}(j)_i} + \frac{2}{3} \sum_{k=1}^{p} r_{X_k(j)_i} r_{X_{\mathbf{k}}(j)_i} \right) \alpha_k + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(-2 \sum_{k=1}^{p} c_{X_k(j)_{n-i+1}} c_{X_{\mathbf{k}}(j)_i} + \frac{2}{3} \sum_{k=1}^{p} r_{X_k(j)_{n-i+1}} r_{X_{\mathbf{k}}(j)_i} \right) \beta_k + \\
&\quad + \sum_{j=1}^{m}\sum_{i=1}^{n} 2 p_i c_{X_{\mathbf{k}}(j)_i} \gamma + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(-2 c_{Y(j)_i} c_{X_{\mathbf{k}}(j)_i} - \frac{2}{3} r_{Y(j)_i} r_{X_{\mathbf{k}}(j)_i} \right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial SE(B)}{\partial \beta_{\mathbf{k}}} &= \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left[2 \left(c_{Y(j)_i} - \gamma - \sum_{k=1}^{p} \left(\alpha_k c_{X_k(j)_i} + \beta_k \left(-c_{X_k(j)_{n-i+1}} \right) \right) \right) \left(c_{X_{\mathbf{k}}(j)_{n-i+1}} \right) + \frac{2}{3} \left(r_{Y(j)_i} - \sum_{k=1}^{p} \left(\alpha_k r_{X_k(j)_i} + \beta_k r_{X_k(j)_{n-i+1}} \right) \right) \left(-r_{X_{\mathbf{k}}(j)_{n-i+1}} \right) \right] = \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(-2 \sum_{k=1}^{p} c_{X_k(j)_i} c_{X_{\mathbf{k}}(j)_{n-i+1}} + \frac{2}{3} \sum_{k=1}^{p} r_{X_k(j)_i} r_{X_{\mathbf{k}}(j)_{n-i+1}} \right) \alpha_k + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(2 \sum_{k=1}^{p} c_{X_k(j)_{n-i+1}} c_{X_{\mathbf{k}}(j)_{n-i+1}} + \frac{2}{3} \sum_{k=1}^{p} r_{X_k(j)_{n-i+1}} r_{X_{\mathbf{k}}(j)_{n-i+1}} \right) \beta_k + \\
&\quad + \sum_{j=1}^{m}\sum_{i=1}^{n} -2 p_i c_{X_{\mathbf{k}}(j)_{n-i+1}} \gamma + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(2 c_{Y(j)_i} c_{X_{\mathbf{k}}(j)_{n-i+1}} - \frac{2}{3} r_{Y(j)_i} r_{X_{\mathbf{k}}(j)_{n-i+1}} \right)
\end{aligned}$$

$$\begin{aligned}
\frac{\partial SE(B)}{\partial \gamma} &= \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left[-2 \left(c_{Y(j)_i} - \sum_{k=1}^{p} \left(\alpha_k c_{X_k(j)_i} + \beta_k \left(-c_{X_k(j)_{n-i+1}} \right) \right) - \gamma \right) \right] = \\
&= \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(2 \sum_{k=1}^{p} \alpha_k c_{X_k(j)_i} \right) + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(-2 \sum_{k=1}^{p} \beta_k c_{X_k(j)_{n-i+1}} \right) + 2m\gamma - \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(2 c_{Y(j)_i} \right)
\end{aligned}$$

Appendix B: Proof of Property 3.5.

Defining the quantile functions $\Psi^{-1}_{Y(j)}(t)$ and $\Psi^{-1}_{\widehat{Y}(j)}(t)$ from the centers and half-ranges of the subintervals, according 3, in *Section 2.1* we have,

$$
\begin{aligned}
&\sum_{j=1}^{m}\int_{0}^{1}\left(\Psi^{-1}_{Y(j)}(t)-\Psi^{-1}_{\widehat{Y}(j)}(t)\right)\left(\Psi^{-1}_{\widehat{Y}(j)}(t)-\overline{Y}\right)dt = \\
&=\sum_{j=1}^{m}\sum_{i=1}^{n}\int_{w_{i-1}}^{w_i}\left[c_{Y(j)_i}+\left(2\frac{t-w_{i-1}}{w_i-w_{i-1}}-1\right)r_{Y(j)_i}-c_{\widehat{Y}(j)_i}-\left(2\frac{t-w_{i-1}}{w_i-w_{i-1}}-1\right)r_{\widehat{Y}(j)_i}\right]\left[c_{\widehat{Y}(j)_i}+\left(2\frac{t-w_{i-1}}{w_i-w_{i-1}}-1\right)r_{\widehat{Y}(j)_i}-\overline{Y}\right]dt = \\
&=\sum_{j=1}^{m}\sum_{i=1}^{n}\int_{w_{i-1}}^{w_i}\left[\left(c_{Y(j)_i}-c_{\widehat{Y}(j)_i}\right)+\left(r_{Y(j)_i}-r_{\widehat{Y}(j)_i}\right)\left(2\frac{t-w_{i-1}}{w_i-w_{i-1}}-1\right)\right]\left[\left(c_{\widehat{Y}(j)_i}-\overline{Y}\right)+r_{\widehat{Y}(j)_i}\left(2\frac{t-w_{i-1}}{w_i-w_{i-1}}-1\right)\right]dt = \\
&=\sum_{j=1}^{m}\sum_{i=1}^{n}\int_{w_{i-1}}^{w_i}\left(c_{Y(j)_i}-c_{\widehat{Y}(j)_i}\right)\left(c_{\widehat{Y}(j)_i}-\overline{Y}\right)+\left(\left(c_{Y(j)_i}-c_{\widehat{Y}(j)_i}\right)r_{\widehat{Y}(j)_i}+\left(r_{Y(j)_i}-r_{\widehat{Y}(j)_i}\right)\left(c_{\widehat{Y}(j)_i}-\overline{Y}\right)\right)\left(2\frac{t-w_{i-1}}{w_i-w_{i-1}}-1\right)+\left(r_{Y(j)_i}-r_{\widehat{Y}(j)_i}\right)r_{\widehat{Y}(j)_i}\left(2\frac{t-w_{i-1}}{w_i-w_{i-1}}-1\right)^2dt
\end{aligned}
$$

Solving the definite integral, after some algebra and considering $w_i - w_{i-1} = p_i$, we obtain,

$$
\sum_{j=1}^{m}\sum_{i=1}^{n}p_i\left[\left(c_{Y(j)_i}-c_{\widehat{Y}(j)_i}\right)c_{\widehat{Y}(j)_i}+\frac{1}{3}\left(r_{Y(j)_i}-r_{\widehat{Y}(j)_i}\right)r_{\widehat{Y}(j)_i}-\left(c_{Y(j)_i}-c_{\widehat{Y}(j)_i}\right)\overline{Y}\right]
$$

For the *equation 13* in *Section 3.2*, $c_{\widehat{Y}(j)_i}=\sum_{k=1}^{p}\alpha_k^* c_{X_k(j)_i}-\beta_k^* c_{X_k(j)_{n-i+1}}+\gamma^*$; $r_{\widehat{Y}(j)_i}=\sum_{k=1}^{p}\alpha_k^* r_{X_k(j)_i}+\beta_k^* r_{X_k(j)_{n-i+1}}$ and for the *Property 3.3* also in *Section 3.2* we have $\overline{Y}=\overline{\widehat{Y}}$, so

$$\sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left[\left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i} \right) \left(\sum_{k=1}^{p} \alpha_k^* c_{X_k(j)_i} - \beta_k^* c_{X_k(j)_{n-i+1}} + \gamma^* \right) + \frac{1}{3} \left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i} \right) \left(\sum_{k=1}^{p} \alpha_k^* r_{X(j)_i} + \beta_k^* r_{X(j)_{n-i+1}} \right) - \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i} \right) \overline{\widehat{Y}} \right] =$$

$$= \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left[\left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i} \right) \sum_{k=1}^{p} \alpha_k^* c_{X_k(j)_i} + \frac{1}{3} \left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i} \right) \sum_{k=1}^{p} \alpha_k^* r_{X_k(j)_i} \right] + \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left[- \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i} \right) \sum_{k=1}^{p} \beta_k^* c_{X_k(j)_{n-i+1}} + \frac{1}{3} \left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i} \right) \sum_{k=1}^{p} \beta_k^* r_{X_k(j)_{n-i+1}} \right] +$$

$$+ \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i} \right) \left(\gamma^* - \overline{\widehat{Y}} \right) =$$

$$= \sum_{j=1}^{m}\sum_{i=1}^{n} \alpha_1^* \left[p_i \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i} \right) c_{X_1(j)_i} + \frac{1}{3} \left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i} \right) r_{X_1(j)_i} \right] + \ldots + \alpha_p^* \left[p_i \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i} \right) c_{X_p(j)_i} + \frac{1}{3} \left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i} \right) r_{X_p(j)_i} \right] +$$

$$+ \sum_{j=1}^{m}\sum_{i=1}^{n} \beta_1^* \left[p_i \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i} \right) \left(-c_{X_1(j)_{n-i+1}} \right) + \frac{1}{3} \left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i} \right) r_{X_1(j)_{n-i+1}} \right] + \ldots + \alpha_p^* \left[p_i \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i} \right) \left(-c_{X_p(j)_{n-i+1}} \right) + \frac{1}{3} \left(r_{Y(j)_i} - r_{\widehat{Y}(j)_i} \right) r_{X_p(j)_{n-i+1}} \right] -$$

$$- \sum_{j=1}^{m}\sum_{i=1}^{n} p_i \left(c_{Y(j)_i} - c_{\widehat{Y}(j)_i} \right) \left(\gamma^* - \overline{\widehat{Y}} \right)$$

Comparing this expression with the partial derivatives of the function SE (see *Appendix A*) we may write

$$\sum_{j=1}^{m} \int_0^1 \left(\Psi_{Y(j)}^{-1}(t) - \Psi_{\widehat{Y}(j)}^{-1}(t) \right) \left(\Psi_{\widehat{Y}(j)}^{-1}(t) - \overline{Y} \right) dt = -\frac{1}{2} \sum_{k=1}^{p} \alpha_k^* \frac{\partial SE(B^*)}{\partial \alpha_k} - \frac{1}{2} \sum_{k=1}^{p} \beta_k^* \frac{\partial SE(B^*)}{\partial \beta_k} + \frac{1}{2} \sum_{k=1}^{p} \frac{\partial SE(B^*)}{\partial \gamma} \left(\gamma^* - \overline{\widehat{Y}} \right)$$

From the Kuhn Tucker conditions presented in *Section 3.3*, we have $\frac{\partial SE(B^*)}{\partial \gamma} = 0$; $\frac{\partial SE(B^*)}{\partial \alpha_k} \alpha_k^* = 0$ and $\frac{\partial SE(B^*)}{\partial \beta_k} \beta_k^* = 0$ for all $k \in \{1, \ldots, p\}$ and $B^* = \left[\alpha_1^* \quad \beta_1^* \quad \alpha_2^* \quad \beta_2^* \quad \cdots \quad \alpha_n^* \quad \beta_n^* \quad \gamma^* \right]^T$. So,

$$\sum_{j=1}^{m} \int_0^1 \left(\Psi_{Y(j)}^{-1}(t) - \Psi_{\widehat{Y}(j)}^{-1}(t) \right) \left(\Psi_{\widehat{Y}(j)}^{-1}(t) - \overline{Y} \right) dt = 0. \qquad \square$$

Appendix C: Simulation results of the study presented in *Subsection 4.1*.

Parameters	*Degree of linearity*	n	*Parameters estimated*						*Goodness of fit measures*			
			$\overline{\alpha}\ (s)$	MSE(α)	$\overline{\beta}\ (s)$	MSE(β)	$\overline{\gamma}\ (s)$	MSE(γ)	$\overline{\Omega}\ (s)$	$\overline{RMSE}_M\ (s)$	$\overline{RMSE}_L\ (s)$	$\overline{RMSE}_U\ (s)$
$\alpha = 2$ $\beta = 1$ $\gamma = -1$	High linearity	10	1.9986 (0.1734)	0.0300	1.0000 (0.1740)	0.0303	−0.9955 (0.1208)	0.0146	0.9741 (0.0089)	0.3450 (0.0628)	0.2985 (0.0585)	0.3001 (0.0592)
		30	2.0033 (0.0881)	0.0078	0.9975 (0.0871)	0.0076	−0.9999 (0.0824)	0.0068	0.9633 (0.0065)	0.3977 (0.0362)	0.3448 (0.0367)	0.3464 (0.0372)
		100	1.9995 (0.0460)	0.0021	1.0008 (0.0464)	0.0021	−0.9994 (0.0354)	0.0013	0.9648 (0.0032)	0.3645 (0.0172)	0.3176 (0.0176)	0.3191 (0.0179)
		250	2.0000 (0.0274)	7.4928E-4	1.0000 (0.0273)	7.4639E-4	−0.9995 (0.0232)	5.3692E-4	0.9640 (0.0020)	0.3604 (0.0105)	0.3145 (0.0110)	0.3160 (0.0111)
	Moderate linearity	10	2.0112 (0.6726)	0.4520	0.9990 (0.6534)	0.4266	−0.9971 (0.4725)	0.2230	0.7163 (0.0786)	1.3821 (0.2527)	1.1986 (0.2427)	1.2048 (0.2461)
		30	1.9832 (0.3517)	0.1239	1.0161 (0.3517)	0.1231	−1.0032 (0.3233)	0.1044	0.6266 (0.0478)	1.5915 (0.1500)	1.3833 (0.1515)	1.3897 (0.1531)
		100	1.9938 (0.1920)	0.0369	1.0065 (0.1909)	0.0364	−1.0020 (0.1461)	0.0213	0.6332 (0.0238)	1.4580 (0.0682)	1.2703 (0.0708)	1.2765 (0.0718)
		250	1.9968 (0.1074)	0.0115	1.0035 (0.1078)	0.0116	−0.9989 (0.0892)	0.0079	0.6269 (0.0156)	1.4399 (0.0444)	1.2559 (0.0455)	1.2618 (0.0461)
	Low linearity	10	1.9796 (1.0974)	1.2035	1.1114 (1.0461)	1.1057	−1.0057 (1.0205)	1.0404	0.4211 (0.1120)	2.7774 (0.5310)	2.4158 (0.5065)	2.4317 (0.5141)
		30	2.0032 (0.7006)	0.4903	1.0160 (0.6671)	0.4448	−1.0200 (0.6091)	0.3710	0.3116 (0.0540)	3.1793 (0.2838)	2.7653 (0.2920)	2.7774 (0.2958)
		100	2.0131 (0.3738)	0.1397	0.9873 (0.3676)	0.1352	−1.0003 (0.2974)	0.0884	0.3054 (0.0268)	2.9212 (0.1417)	2.5476 (0.1448)	2.5604 (0.1467)
		250	2.0005 (0.2148)	0.0461	0.9995 (0.2158)	0.0465	−1.0045 (0.1799)	0.0323	0.2969 (0.0167)	2.8845 (0.0831)	2.5179 (0.0865)	2.5301 (0.0879)
$\alpha = 2$ $\beta = 8$ $\gamma = 3$	High linearity	10	2.2019 (0.5499)	0.3026	7.9802 (0.5503)	0.3029	2.9887 (0.3987)	0.1589	0.9761 (0.0083)	1.1381 (0.2097)	0.9868 (0.1975)	0.9933 (0.1996)
		30	1.9903 (0.3168)	0.1004	8.0103 (0.3152)	0.0993	2.9904 (0.2642)	0.0698	0.9663 (0.0061)	1.3260 (0.1220)	1.1526 (0.1224)	1.1573 (0.1238)
		100	2.0040 (0.1533)	0.0235	7.9962 (0.1539)	0.0237	2.9970 (0.1138)	0.0129	0.9691 (0.0030)	1.1871 (0.0580)	1.0348 (0.0602)	1.0400 (0.0609)
		250	1.9999 (0.0848)	0.0072	7.9995 (0.0855)	0.0073	2.9982 (0.0713)	0.0051	0.9692 (0.0018)	1.1689 (0.0346)	1.0198 (0.0358)	1.10250 (0.0364)
	Moderate linearity	10	2.2604 (1.9509)	3.8700	7.8251 (2.0929)	4.4065	2.9568 (1.5889)	2.2538	0.7264 (0.0767)	4.5927 (0.8567)	3.9892 (0.8203)	4.0107 (0.8335)
		30	1.9800 (1.1222)	1.2584	8.0427 (1.1444)	1.3102	2.9721 (1.0640)	1.1317	0.6443 (0.0472)	5.3435 (0.4927)	4.6441 (0.5078)	4.6669 (0.5117)
		100	1.9740 (0.6063)	0.3679	8.0270 (0.6062)	0.3678	3.0167 (0.4742)	0.2249	0.6630 (0.0254)	4.7626 (0.2319)	4.1546 (0.2403)	4.1759 (0.2433)
		250	2.0086 (0.3528)	0.1244	7.9893 (0.3542)	0.1254	3.0101 (0.2964)	0.0879	0.6627 (0.0154)	4.6817 (0.1378)	4.0839 (0.1421)	4.1048 (0.1441)
	Low linearity	10	2.9698 (3.2472)	11.4742	7.4940 (3.7303)	14.1576	2.7300 (3.0986)	9.6644	0.4347 (0.1062)	9.1566 (1.6951)	7.9664 (1.6167)	8.0159 (1.6382)
		30	2.2606 (1.9869)	4.0117	7.8835 (2.2006)	4.8515	3.0435 (2.0941)	4.3826	0.3230 (0.0588)	10.6817 (1.0166)	9.3053 (1.0340)	9.3463 (1.0485)
		100	2.0660 (1.1773)	1.3890	7.9436 (1.2087)	1.4627	3.0111 (0.9445)	0.8912	0.3308 (0.0321)	9.5497 (0.4346)	8.3331 (0.4467)	8.3765 (0.4536)
		250	1.9750 (0.6901)	0.4764	8.0194 (0.6930)	0.4801	3.0063 (0.5768)	0.3324	0.3311 (0.0211)	9.3675 (0.2735)	8.1738 (0.2793)	8.2163 (0.2833)
$\alpha = 8$ $\beta = 0$ $\gamma = 4$	High linearity	10	7.8819 (0.3867)	0.1633	0.2075 (0.3046)	0.1357	4.0043 (0.3542)	0.1254	0.9705 (0.0094)	1.0863 (0.1895)	0.9491 (0.1864)	0.9533 (0.1877)
		30	7.9635 (0.1817)	0.0343	0.0918 (0.1309)	0.0255	3.9800 (0.1947)	0.0383	0.9734 (0.0046)	1.0348 (0.0935)	0.9023 (0.0960)	0.9063 (0.0972)
		100	7.9735 (0.1166)	0.0143	0.0583 (0.0852)	0.0106	4.0041 (0.1209)	0.0146	0.9626 (0.0036)	1.1581 (0.0558)	1.0079 (0.0583)	1.0115 (0.0590)
		250	7.9897 (0.0698)	0.0050	0.0326 (0.0478)	0.0033	4.0023 (0.0675)	0.0046	0.9654 (0.0021)	1.1178 (0.0339)	0.9744 (0.0352)	0.9777 (0.0357)
	Moderate linearity	10	7.3294 (1.5775)	2.9356	0.9722 (1.3090)	2.6569	4.0382 (1.4967)	2.2395	0.6868 (0.0799)	4.2833 (0.7995)	3.7237 (0.7780)	3.7410 (0.7834)
		30	7.8624 (0.7205)	0.5376	0.3507 (0.5251)	0.3984	3.9431 (0.7535)	0.5705	0.6967 (0.0440)	4.1547 (0.3757)	3.6279 (0.3806)	3.6457 (0.3865)
		100	7.8926 (0.4713)	0.2334	0.2370 (0.3470)	0.1765	4.0522 (0.4661)	0.2197	0.6179 (0.0290)	4.6283 (0.2173)	4.0338 (0.2267)	4.0481 (0.2294)
		250	7.9589 (0.2733)	0.0763	0.1289 (0.1876)	0.0518	4.0069 (0.2743)	0.0752	0.6362 (0.0186)	4.4661 (0.1328)	3.8956 (0.1417)	3.9090 (0.1430)
	Low linearity	10	7.0359 (2.9315)	9.5147	1.6765 (2.4074)	8.6006	4.1706 (2.8774)	8.3002	0.3802 (0.1072)	8.6643 (1.5792)	7.5591 (1.5361)	7.5892 (1.5503)
		30	7.5575 (1.4950)	2.4286	0.8096 (1.1388)	1.9510	3.8491 (1.5347)	2.3757	0.3690 (0.0643)	8.2692 (0.7530)	7.2169 (0.7613)	7.2493 (0.7713)
		100	7.7974 (0.9858)	1.0118	0.4802 (0.7177)	0.7452	4.0602 (0.9472)	0.8999	0.2907 (0.0368)	9.2602 (0.4236)	8.0586 (0.4474)	8.0864 (0.4505)
		250	7.8938 (0.5627)	0.3276	0.2748 (0.4009)	0.2361	4.0053 (0.5556)	0.3084	0.3038 (0.0250)	8.9426 (0.2601)	7.7952 (0.2758)	7.8224 (0.2788)

Table 6: Results of the *DSD Model* when the histogram values of X are generated from real values with an Uniform distribution.

Parameters	*Degree of linearity*	n	*Parameters estimated*						*Goodness of fit measures*			
			$\overline{\widehat{\alpha}}\,(s)$	MSE(α)	$\overline{\widehat{\beta}}\,(s)$	MSE(β)	$\overline{\widehat{\gamma}}\,(s)$	MSE(γ)	$\overline{\Omega}\,(s)$	$\overline{RMSE}_M\,(s)$	$\overline{RMSE}_L\,(s)$	$\overline{RMSE}_U\,(s)$
$\alpha = 2$ $\beta = 1$ $\gamma = -1$	High linearity	10	2.0178 (0.3322)	0.1106	0.9834 (0.3333)	0.1112	−1.0148 (0.3809)	0.1451	0.9792 (0.0079)	0.5101 (0.1002)	0.4470 (0.0879)	0.4602 (0.0918)
		30	2.0064 (0.1967)	0.0387	0.9931 (0.1967)	0.0387	−1.0113 (0.2222)	0.0486	0.9757 (0.0047)	0.6738 (0.0669)	0.5868 (0.0603)	0.6051 (0.0638)
		100	1.9977 (0.1116)	0.0124	1.0021 (0.1118)	0.0125	−0.9980 (0.1298)	0.0168	0.9762 (0.0025)	0.6720 (0.0359)	0.5819 (0.0322)	0.6002 (0.0341)
		250	1.9958 (0.0728)	0.0053	1.0046 (0.0725)	0.0053	−0.9948 (0.0874)	0.0077	0.9765 (0.0016)	0.6635 (0.0227)	0.5744 (0.0204)	0.5931 (0.0219)
	Moderate linearity	10	1.8845 (1.0124)	1.0373	1.1353 (0.9992)	1.0157	−0.8845 (1.1579)	1.3528	0.7567 (0.0691)	2.0408 (0.3787)	1.7920 (0.3352)	1.8408 (0.3525)
		30	1.9496 (0.7194)	0.5201	1.0616 (0.7140)	0.5130	−0.9506 (0.8169)	0.6691	0.7202 (0.0409)	2.6920 (0.2632)	2.3416 (0.2345)	2.4114 (0.2498)
		100	2.0133 (0.4457)	0.1986	0.9857 (0.4447)	0.1978	−1.0186 (0.5086)	0.2588	0.7217 (0.0232)	2.6818 (0.1477)	2.3247 (0.1325)	2.3958 (0.1409)
		250	1.9926 (0.2746)	0.0754	1.0096 (0.2734)	0.0747	−0.9818 (0.3240)	0.1052	0.7230 (0.0138)	2.6544 (0.0885)	2.2978 (0.0797)	2.3727 (0.0844)
	Low linearity	10	1.7251 (1.3442)	1.8805	1.3722 (1.3365)	1.9230	−0.6851 (1.8487)	3.5134	0.4587 (0.1013)	4.1109 (0.7661)	3.6144 (0.6786)	3.7240 (0.7154)
		30	1.8106 (1.0925)	1.2283	1.2173 (1.0791)	1.2105	−0.8098 (1.4359)	2.0959	0.3991 (0.0533)	5.3963 (0.5169)	4.6922 (0.4582)	4.8332 (0.4888)
		100	1.9243 (0.7792)	0.6123	1.0779 (0.7731)	0.6031	−0.9353 (0.9275)	0.8636	0.3970 (0.0306)	5.3543 (0.2923)	4.6376 (0.2592)	4.7804 (0.2737)
		250	1.9894 (0.5271)	0.2777	1.0098 (0.5258)	0.2763	−1.0040 (0.6464)	0.4174	0.3954 (0.0180)	5.3114 (0.1752)	4.5970 (0.1557)	4.7446 (0.1656)
$\alpha = 2$ $\beta = 8$ $\gamma = 3$	High linearity	10	2.0375 (1.0727)	1.1509	7.9640 (1.0767)	1.1594	2.9425 (1.2111)	1.4687	0.9802 (0.0075)	1.6645 (0.3292)	1.4621 (0.2869)	1.5004 (0.3006)
		30	2.0435 (0.6690)	0.4490	7.9542 (0.6707)	0.4515	2.9695 (0.7301)	0.5335	0.9766 (0.0047)	2.2147 (0.2266)	1.9273 (0.2072)	1.9820 (0.2158)
		100	1.9892 (0.3691)	0.1362	8.0112 (0.3700)	0.1369	3.0199 (0.4205)	0.1770	0.9769 (0.0024)	2.2171 (0.1188)	1.9225 (0.1081)	1.9844 (0.1143)
		250	1.9985 (0.2246)	0.0504	8.0009 (0.2255)	0.0508	3.0013 (0.2713)	0.0735	0.9770 (0.0016)	2.2006 (0.0763)	1.9058 (0.0675)	1.9651 (0.0728)
	Moderate linearity	10	2.8185 (3.0480)	9.9508	7.2759 (3.1530)	10.4560	2.2950 (3.8358)	15.1956	0.7625 (0.0714)	6.7185 (1.3172)	5.9011 (1.1491)	6.0540 (1.2196)
		30	2.3178 (2.1781)	4.8406	7.7115 (2.2241)	5.0248	2.6605 (2.5909)	6.8212	0.7253 (0.0411)	8.8999 (0.8947)	7.7497 (0.7979)	7.9810 (0.8575)
		100	2.0196 (1.3285)	1.7634	7.9920 (1.3353)	1.7814	2.9729 (1.5781)	2.4885	0.7279 (0.0221)	8.8561 (0.4802)	7.6712 (0.4317)	7.9220 (0.4560)
		250	1.9710 (0.8743)	0.7645	8.0308 (0.8762)	0.7678	3.0363 (1.0483)	1.0991	0.7269 (0.0139)	8.8048 (0.2916)	7.6183 (0.2597)	7.8600 (0.2694)
	Low linearity	10	3.6276 (4.1434)	19.8002	6.6974 (4.2852)	20.0415	1.2700 (6.1575)	40.8694	0.4552 (0.1021)	13.7941 (2.5462)	12.1258 (2.2524)	12.4544 (2.3568)
		30	3.2768 (3.4625)	13.6072	6.8516 (3.5649)	14.0145	1.9774 (4.5906)	22.0986	0.4089 (0.0569)	17.7236 (1.8438)	15.4633 (1.6399)	15.9189 (1.7256)
		100	2.4832 (2.3068)	5.5496	7.5543 (2.3403)	5.6700	2.5576 (2.8965)	8.5772	0.4030 (0.0281)	17.7182 (0.9047)	15.3586 (0.8263)	15.8561 (0.8678)
		250	2.0307 (1.5349)	2.3546	7.9699 (1.5551)	2.4169	2.9507 (1.9185)	3.6793	0.3992 (0.0194)	17.6535 (0.6215)	15.2807 (0.5501)	15.7680 (0.5869)
$\alpha = 8$ $\beta = 0$ $\gamma = 4$	High linearity	10	7.5241 (0.7522)	0.7918	0.5138 (0.7162)	0.7765	4.4600 (0.9771)	1.1653	0.9584 (0.0146)	1.9865 (0.3801)	1.7382 (0.3559)	1.7580 (0.3630)
		30	7.7331 (0.4338)	0.2592	0.2894 (0.4152)	0.2560	4.2539 (0.5755)	0.3953	0.9626 (0.0068)	2.2925 (0.2204)	1.9940 (0.2083)	2.0307 (0.2184)
		100	7.8596 (0.2314)	0.0732	0.1526 (0.2217)	0.0732	4.1456 (0.3186)	0.1226	0.9609 (0.0038)	2.3586 (0.1184)	2.0515 (0.1152)	2.0848 (0.1203)
		250	7.9109 (0.1474)	0.0297	0.0969 (0.1419)	0.0295	4.0954 (0.2074)	0.0521	0.9600 (0.0024)	2.3755 (0.0739)	2.0661 (0.0714)	2.1029 (0.0742)
	Moderate linearity	10	6.1982 (2.8443)	11.3282	2.0033 (2.7031)	11.3125	5.8740 (3.9169)	18.8386	0.6074 (0.0880)	7.9771 (1.4476)	6.9641 (1.3661)	7.0623 (1.4146)
		30	6.9001 (1.8034)	4.4586	1.1966 (1.7298)	4.4213	5.1608 (2.3103)	4.4213	0.6223 (0.0455)	9.1543 (0.8663)	7.9724 (0.8349)	8.1157 (0.8628)
		100	7.4109 (0.9902)	1.3265	0.6393 (0.9509)	1.3120	4.6070 (1.3341)	2.1464	0.6067 (0.0241)	9.4459 (0.4553)	8.2107 (0.4397)	8.3466 (0.4580)
		250	7.6613 (0.5593)	0.4273	0.3719 (0.5307)	0.4196	4.3825 (0.8076)	0.7978	0.6004 (0.0155)	9.5071 (0.2960)	8.2721 (0.2938)	8.4202 (0.3019)
	Low linearity	10	5.7584 (3.7482)	19.0597	2.7877 (3.5915)	20.6570	6.3648 (6.4528)	47.1898	0.3029 (0.0891)	16.0701 (2.9634)	14.0612 (2.8037)	14.2383 (2.8682)
		30	6.0472 (3.0115)	12.8732	2.1679 (2.8934)	13.0634	5.8323 (4.3960)	22.6630	0.2976 (0.0458)	18.4440 (1.7816)	16.0750 (1.7212v	16.3662 (1.7986)
		100	6.8326 (1.8938)	4.956	1.2632 (1.8194)	4.9026	5.2975 (2.6368)	8.6293	0.2806 (0.0248)	18.8923 (0.9458)	16.4375 (0.9135)	16.7107 (0.9502)
		250	7.2406 (1.2367)	2.1047	0.8182 (1.1878)	2.0788	4.8289 (1.6966)	3.5626	0.2736 (0.0162)	19.0168 (0.6280)	16.5351 (0.6009)	16.8232 (0.6316)

Table 7: Results of the *DSD Model* when the histogram values of X are generated from real values with a Normal distribution.

Parameters	*Degree of linearity*	n	*Parameters estimated*						*Goodness of fit measures*			
			$\overline{\widehat{\alpha}}\,(s)$	MSE(α)	$\overline{\widehat{\beta}}\,(s)$	MSE(β)	$\overline{\widehat{\gamma}}\,(s)$	MSE(γ)	$\overline{\Omega}\,(s)$	$\overline{RMSE}_M\,(s)$	$\overline{RMSE}_L\,(s)$	$\overline{RMSE}_U\,(s)$
$\alpha = 2$ $\beta = 1$ $\gamma = -1$	High linearity	10	1.9998 (0.0581)	0.0034	1.0002 (0.0229)	5.2466E-4	−1.0104 (0.5485)	0.3007	0.9843 (0.0054)	1.7361 (0.3009)	1.4034 (0.2684)	1.7717 (0.2920)
		30	1.9979 (0.0338)	0.0011	1.0006 (0.0141)	1.9756E-4	−0.9938 (0.2260)	0.0511	0.9821 (0.0034)	1.2068 (0.1155)	0.9828 (0.1020)	1.2178 (0.1113)
		100	2.0001 (0.0233)	5.4282E-4	1.0002 (0.0080)	6.4532E-5	−1.0004 (0.1154)	0.0133	0.9822 (0.0019)	1.1460 (0.0615)	0.9202 (0.0527)	1.1344 (0.0561)
		250	2.0003 (0.0168)	2.8347E-4	0.9998 (0.0050)	2.5440E-5	−1.0005 (0.0813)	0.0066	0.9828 (0.0012)	1.3450 (0.0472)	1.0631 (0.0384)	1.3249 (0.0418)
	Moderate linearity	10	2.0100 (0.2298)	0.0528	1.0000 (0.0966)	0.0093	−1.0167 (2.1989)	4.8305	0.8032 (0.0587)	6.8729 (1.1688)	5.5639 (1.0222)	7.0662 (1.1728)
		30	1.9986 (0.1396)	0.0195	0.9996 (0.0554)	0.0031	−1.0145 (0.8822)	0.7778	0.7735 (0.0376)	4.8442 (0.4605)	3.9500 (0.4087)	4.8851 (0.4433)
		100	1.9980 (0.0947)	0.0090	1.0017 (0.0334)	0.0011	−0.9844 (0.4405)	0.1941	0.7765 (0.0229)	4.5678 (0.2531)	3.6677 (0.2139)	4.5260 (0.2326)
		250	1.9994 (0.0628)	0.0039	0.9994 (0.0199)	3.9465E-4	−1.0030 (0.3278)	0.1074	0.7808 (0.0142)	5.3830 (0.1813)	4.2525 (0.1491)	5.2980 (0.1625)
	Low linearity	10	1.9966 (0.4707)	0.2214	0.9932 (0.1924)	0.0370	−1.0635 (4.4255)	19.5691	0.5067 (0.1208)	13.8276 (2.4178)	11.2257 (2.1118)	14.2114 (2.3597)
		30	2.0063 (0.2585)	0.0668	0.9990 (0.1168)	0.0136	−1.0622 (1.7812)	3.1732	0.4624 (0.0652)	9.7396 (0.9122)	7.9333 (0.8016)	9.8241 (0.8739)
		100	1.9945 (0.1857)	0.0345	1.0002 (0.0669)	0.0045	−0.9811 (0.8548)	0.7304	0.4638 (0.0418)	9.1568 (0.4889)	7.3582 (0.4072)	9.0718 (0.4576)
		250	1.9999 (0.1301)	0.0169	1.0029 (0.0390)	0.0015	−0.9549 (0.6359)	0.4060	0.4719 (0.0291)	10.7653 (0.3713)	8.5066 (0.2978)	10.6124 (0.3263)
$\alpha = 2$ $\beta = 8$ $\gamma = 3$	High linearity	10	1.9959 (0.1147)	0.0132	7.9975 (0.1421)	0.0202	2.9200 (2.9243)	8.5494	0.9697 (0.0123)	8.7007 (1.8586)	7.4852 (1.4434)	7.3627 (1.5829)
		30	2.0014 (0.0638)	0.0041	8.0023 (0.0894)	0.0080	3.0171 (1.1609)	1.3466	0.9671 (0.0075)	5.8900 (0.6950)	5.0448 (0.5458)	4.9153 (0.6076)
		100	1.9980 (0.0589)	0.0035	8.0008 (0.0426)	0.0018	3.0090 (0.5439)	0.2956	0.9674 (0.0047)	5.5953 (0.4134)	4.6336 (0.2904)	4.5220 (0.3222)
		250	1.9984 (0.0411)	0.0017	8.0001 (0.0261)	6.7870E-4	3.0078 (0.3997)	0.1597	0.9668 (0.0031)	6.7720 (0.3279)	5.4657 (0.2144)	5.3485 (0.2358)
	Moderate linearity	10	2.0193 (0.4599)	0.2117	8.0077 (0.5951)	0.3539	3.5144 (11.9446)	142.7964	0.6753 (0.0946)	34.9590 (7.3023)	30.0676 (5.7685)	29.6291 (6.3318)
		30	2.0162 (0.2589)	0.0672	7.9990 (0.3651)	0.1332	2.9598 (4.5878)	21.0289	0.6501 (0.0581)	23.6253 (2.8726)	20.2518 (2.2448)	19.7469 (2.5017)
		100	1.9969 (0.2429)	0.0590	7.9937 (0.1647)	0.0271	2.9606 (2.1746)	4.7259	0.6488 (0.0330)	22.4810 (1.5323)	18.5825 (1.0644)	18.1421 (1.1779)
		250	1.9924 (0.1627)	0.0265	7.9999 (0.1109)	0.0123	2.9652 (1.6590)	2.7508	0.6461 (0.0228)	27.0579 (1.2875)	21.8555 (0.8409)	21.3869 (0.9177)
	Low linearity	10	2.0207 (0.9350)	0.8737	7.9917 (1.1739)	1.3767	3.4413 (22.9672)	527.1604	0.3516 (0.1213)	70.8050 (15.1479)	60.6901 (12.0727)	59.9225 (13.1667)
		30	2.0255 (0.5530)	0.3061	8.0094 (0.6928)	0.4796	3.1185 (9.0410)	81.6715	0.3229 (0.0637)	47.2017 (5.5543)	40.4264 (4.2789)	39.4641 (4.7159)
		100	2.0155 (0.4576)	0.2095	8.0133 (0.3395)	0.1153	3.2310 (4.3766)	19.1891	0.3207 (0.0373)	44.8115 (3.2170)	37.0780 (2.2761)	36.1995 (2.4814)
		250	2.0010 (0.3267)	0.1066	7.9960 (0.2142)	0.0458	2.9879 (3.2616)	10.6274	0.3144 (0.0247)	54.0955 (2.5264)	43.6840 (1.6518)	42.7230 (1.7807)
$\alpha = 8$ $\beta = 0$ $\gamma = 4$	High linearity	10	7.9992 (0.2169)	0.0470	0.0237 (0.0351)	0.0018	4.0425 (1.8739)	3.5098	0.9840 (0.0046)	5.9069 (0.8538)	4.5285 (0.9122)	6.3816 (0.9960)
		30	7.9897 (0.1306)	0.0171	0.0204 (0.0278)	0.0012	4.0801 (0.8450)	0.7197	0.9750 (0.0040)	4.7459 (0.3769)	3.7987 (0.4200)	4.9109 (0.4183)
		100	7.9975 (0.0902)	0.0081	0.0138 (0.0202)	5.9900E-4	4.0176 (0.4421)	0.1955	0.9748 (0.0021)	4.5582 (0.1849)	3.6551 (0.2091)	4.6524 (0.2041)
		250	8.0000 (0.0627)	0.0039	0.0075 (0.0109)	1.7389E-4	4.0227 (0.3092)	0.0960	0.9786 (0.0012)	5.0417 (0.1356)	3.9455 (0.1458)	5.2254 (0.1541)
	Moderate linearity	10	7.9502 (0.9056)	0.8218	0.1086 (0.1511)	0.0346	4.4036 (7.2524)	52.7076	0.7873 (0.0580)	23.9256 (3.2503)	18.3651 (3.6504)	25.9095 (3.7127)
		30	7.9880 (0.5306)	0.2814	0.0711 (0.1103)	0.0172	4.1567 (3.3115)	10.9796	0.7132 (0.0399)	18.8067 (1.4343)	15.0325 (1.5704)	19.4762 (1.5704)
		100	7.9813 (0.3595)	0.1294	0.0543 (0.0769)	0.0089	4.1978 (1.7758)	3.1894	0.7074 (0.0243)	18.2054 (0.7060)	14.5693 (0.7915)	18.5745 (0.8071)
		250	7.9967 (0.2429)	0.0589	0.0281 (0.0411)	0.0025	4.0983 (1.2140)	1.4820	0.7412 (0.0158)	20.1525 (0.5361)	15.7578 (0.5703)	20.8938 (0.6275)
	Low linearity	10	7.9569 (1.8656)	3.4790	0.2041 (0.3005)	0.1319	5.0208 (14.4783)	210.4528	0.4873 (0.1287)	47.3893 (6.7317)	36.5172 (7.0611)	51.1697 (7.7885)
		30	7.9415 (1.0892)	1.1887	0.1538 (0.2205)	0.0723	4.4560 (6.5923)	43.6228	0.3807 (0.0705)	37.8052 (2.8646)	30.1898 (3.1327)	39.1813 (3.1481)
		100	7.9739 (0.6931)	0.4806	0.0950 (0.1383)	0.0281	4.1113 (3.6089)	13.0233	0.3765 (0.0436)	36.4753 (1.5110)	29.2075 (1.6813)	37.1965 (1.7049)
		250	8.0034 (0.4882)	0.2381	0.0556 (0.0805)	0.0096	4.1368 (2.5335)	6.4309	0.4173 (0.0314)	40.3673 (1.0888)	31.5864 (1.1539)	41.8406 (1.2501)

Table 8: Results of the *DSD Model* when the histogram values of X are generated from real values with a Log-Normal distribution.

Parameters	*Degree of linearity*	n	*Parameters estimated*						*Goodness of fit measures*			
			$\overline{\widehat{\alpha}}\,(s)$	MSE(α)	$\overline{\widehat{\beta}}\,(s)$	MSE(β)	$\overline{\widehat{\gamma}}\,(s)$	MSE(γ)	$\overline{\Omega}\,(s)$	$\overline{RMSE}_M\,(s)$	$\overline{RMSE}_L\,(s)$	$\overline{RMSE}_U\,(s)$
$\alpha = 2$ $\beta = 1$ $\gamma = -1$	High linearity	10	1.9956 (0.0890)	0.0079	1.0031 (0.0835)	0.0070	−0.9941 (0.2387)	0.0569	0.9780 (0.0078)	0.5395 (0.1001)	0.4720 (0.0908)	0.4909 (0.0945)
		30	2.0000 (0.0411)	0.0017	1.0009 (0.0303)	9.1592E-4	−0.9925 (0.1454)	0.0212	0.9768 (0.0044)	0.7348 (0.0716)	0.6308 (0.0663)	0.6854 (0.0688)
		100	1.995 (0.0347)	0.0012	1.0002 (0.0337)	0.0011	−0.9991 (0.0967)	0.0093	0.9719 (0.0026)	0.5956 (0.0281)	0.5175 (0.0269)	0.5426 (0.0280)
		250	1.9998 (0.0191)	3.6369E-4	1.0003 (0.0169)	2.8659E-4	−0.9985 (0.0502)	0.0025	0.9736 (0.0016)	0.6981 (0.0219)	0.6030 (0.0206)	0.6334 (0.0216)
	Moderate linearity	10	1.9837 (0.3579)	0.1282	1.0135 (0.3320)	0.1103	−0.9892 (0.9200)	0.8457	0.7453 (0.0706)	2.1512 (0.3873)	1.8853 (0.3540)	1.9594 (0.3679)
		30	2.0048 (0.1694)	0.0287	0.9978 (0.1272)	0.0162	−0.9962 (0.5829)	0.3394	0.7274 (0.0400)	2.9372 (0.2721)	2.5176 (0.2509)	2.7405 (0.2667)
		100	2.0014 (0.1482)	0.0219	0.9974 (0.1444)	0.0208	−1.0095 (0.4045)	0.1635	0.6838 (0.0223)	2.3856 (0.1199)	2.0729 (0.1142)	2.1739 (0.1194)
		250	2.0012 (0.0738)	0.0054	0.9993 (0.0671)	0.0045	−0.9960 (0.1996)	0.0398	0.6973 (0.0145)	2.7948 (0.0872)	2.4132 (0.0811)	2.5358 (0.0870)
	Low linearity	10	1.9830 (0.6900)	0.4759	1.0192 (0.6402)	0.4098	−1.0329 (1.8983)	3.6010	0.4422 (0.1024)	4.3516 (0.7826)	3.8048 (0.7121)	3.9528 (0.7438)
		30	1.9956 (0.3416)	0.1166	1.0057 (0.2565)	0.0657	−0.9307 (1.1922)	1.4246	0.4037 (0.0593)	5.8856 (0.5477)	5.0386 (0.5045)	5.4837 (0.5414)
		100	1.9969 (0.2941)	0.0864	1.0025 (0.2834)	0.0803	−0.9855 (0.7845)	0.6151	0.3526 (0.0260)	4.7728 (0.2326)	4.1458 (0.2169)	4.3485 (0.2289)
		250	1.9965 (0.1498)	0.0224	1.0046 (0.1346)	0.0181	−0.9955 (0.4129)	0.1704	0.3668 (0.0193)	5.5841 (0.1766)	4.8207 (0.1615)	5.0676 (0.1756)
$\alpha = 2$ $\beta = 8$ $\gamma = 3$	High linearity	10	2.0006 (0.2809)	0.0788	8.0001 (0.2656)	0.0705	3.0125 (0.7988)	0.6377	0.9756 (0.0099)	2.0152 (0.4220)	1.7458 (0.3323)	1.7582 (0.3583)
		30	1.9987 (0.1011)	0.0102	7.9990 (0.1051)	0.0110	2.9928 (0.5651)	0.3191	0.9699 (0.0069)	2.9473 (0.3548)	2.5058 (0.2713)	2.4971 (0.2955)
		100	2.0020 (0.1102)	0.0121	7.9992 (0.1113)	0.0124	2.9922 (0.3179)	0.1010	0.9678 (0.0035)	2.1769 (0.1227)	1.9067 (0.1085)	1.8959 (0.1187)
		250	2.0001 (0.0574)	0.0033	7.9998 (0.0521)	0.0027	3.0031 (0.1682)	0.0283	0.9719 (0.0022)	2.5514 (0.1024)	2.1680 (0.0783)	2.1988 (0.0834)
	Moderate linearity	10	2.0822 (1.0942)	1.2029	7.9330 (1.0660)	1.1396	2.9400 (3.1397)	9.8516	0.7276 (0.0851)	7.9730 (1.6504)	6.8985 (1.3292)	6.9345 (1.4250)
		30	1.9952 (0.4228)	0.1786	8.0125 (0.4222)	0.1782	3.0396 (2.2061)	4.8637	0.6711 (0.0562)	11.8294 (1.4578)	10.0343 (1.0981)	10.0133 (1.2060)
		100	1.9932 (0.4530)	0.2050	8.0068 (0.4520)	0.2042	3.0202 (1.2916)	1.6670	0.6530 (0.0250)	8.7238 (0.4660)	7.6440 (0.4076)	7.5979 (0.4432)
		250	2.0000 (0.2282)	0.0520	7.9971 (0.2064)	0.0426	3.0142 (0.6866)	0.4712	0.6833 (0.0178)	10.2192 (0.3974)	8.6694 (0.3069)	8.7979 (0.3302)
	Low linearity	10	2.1848 (1.8694)	3.5254	7.9569 (2.0080)	4.0300	2.8979 (6.3806)	40.6812	0.4156 (0.1127)	16.2758 (3.2714)	14.1198 (2.6377)	14.1619 (2.8575)
		30	2.0320 (0.8060)	0.6500	8.0217 (0.8408)	0.7068	3.2052 (4.6480)	21.6242	0.3460 (0.0639)	23.6048 (2.8823)	19.9899 (2.1464)	19.9082 (2.3476)
		100	1.9883 (0.8909)	0.7930	8.0040 (0.8982)	0.8060	2.9768 (2.6469)	6.9995	0.3200 (0.0264)	17.5002 (0.9303)	15.3325 (0.8117)	15.2520 (0.8982)
		250	2.0135 (0.4468)	0.1996	7.9790 (0.4010)	0.1610	2.9814 (1.2923)	1.6686	0.3502 (0.0213)	20.4601 (0.7790)	17.3477 (0.6012)	17.6061 (0.6496)
$\alpha = 8$ $\beta = 0$ $\gamma = 4$	High linearity	10	7.9658 (0.3638)	0.1334	0.1543 (0.2179)	0.0712	4.1461 (0.9313)	0.8877	0.9612 (0.0126)	2.3616 (0.4150)	2.0561 (0.4055)	2.1431 (0.4060)
		30	7.9838 (0.1544)	0.0241	0.0536 (0.0731)	0.0082	4.0908 (0.5604)	0.3219	0.9638 (0.0059)	2.9241 (0.2422)	2.4813 (0.2551)	1.7361 (0.2553)
		100	7.9532 (0.1223)	0.0171	0.0692 (0.0996)	0.0147	4.1273 (0.3734)	0.1555	0.9262 (0.0064)	2.8461 (0.1313)	2.4625 (0.1405)	2.5509 (0.1388)
		250	7.9956 (0.0644)	0.0042	0.0242 (0.377)	0.0020	4.0285 (0.1835)	0.0345	0.9628 (0.0022)	2.6839 (0.0801)	2.3072 (0.0812)	2.4322 (0.0849)
	Moderate linearity	10	7.7845 (1.4164)	2.0506	0.6640 (0.8895)	1.2313	4.8987 (3.8242)	15.4179	0.6133 (0.0955)	9.4563 (1.5639)	8.2210 (1.5300)	8.5593 (1.5483)
		30	7.9183 (0.6171)	0.3871	0.2113 (0.2955)	0.1319	4.3967 (2.3469)	5.6596	0.6246 (0.0486)	11.7120 (1.0040)	9.9480 (1.0538)	10.9854 (1.0670)
		100	7.8282 (0.4778)	0.2576	0.2640 (0.3825)	0.2158	4.5254 (1.4884)	2.4894	0.4421 (0.0267)	11.3698 (0.5352)	9.8308 (0.5707)	10.1877 (0.5654)
		250	7.9784 (0.2569)	0.0664	0.1024 (0.1522)	0.0336	4.1186 (0.7272)	0.5423	0.6179 (0.0191)	10.7424 (0.3234)	9.2335 (0.3343)	9.7364 (0.3505)
	Low linearity	10	7.6509 (2.7407)	7.6257	1.2133 (1.6431)	4.1693	5.7207 (7.6657)	61.6651	0.3064 (0.1092)	18.8776 (3.0931)	16.3992 (3.0835)	17.1068 (3.1387)
		30	7.8026 (1.1907)	1.4552	0.4252 (0.5832)	0.5206	4.7226 (4.4911)	20.6724	0.2967 (0.0619)	23.3571 (1.9443)	19.8547 (2.0551)	21.8660 (2.0591)
		100	7.6032 (0.9440)	1.1474	0.5621 (0.7538)	0.9710	5.1325 (3.0385)	9.8381	0.1655 (0.0200)	22.7438 (1.0629)	19.6830 (1.1269)	20.3798 (1.1215)
		250	7.9611 (0.5330)	0.2853	0.2081 (0.3103)	0.1395	4.2216 (1.4930)	2.2759	0.2897 (0.0256)	21.4657 (0.6328)	18.4560 (0.6494)	19.4450 (0.6683)

Table 9: Results of the *DSD Model* when the histogram values of X are generated from real values from a mixture of different distributions.

Distributions in variables $X_1; X_2; X_3$	*Degree of linearity*	n	*Parameters estimated*											
			$\widehat{\widehat{\alpha_1}}\,(s)$	MSE(α_1)	$\widehat{\widehat{\beta_1}}\,(s)$	MSE(β_1)	$\widehat{\widehat{\alpha_2}}\,(s)$	MSE(α_2)	$\widehat{\widehat{\beta_2}}\,(s)$	MSE(β_2)	$\widehat{\widehat{\alpha_3}}\,(s)$	MSE(α_3)	$\widehat{\widehat{\beta_3}}\,(s)$	MSE(β_3)
Uniform	High linearity	10	1.9790 (0.4625)	0.2141	0.9664 (0.4621)	0.2145	0.5375 (0.4958)	0.2469	3.0505 (0.6273)	0.3957	3.9859 (0.6573)	0.4319	1.9784 (0.6678)	0.4460
		30	2.0005 (0.3101)	0.0961	0.9978 (0.3096)	0.0958	0.4965 (0.2887)	0.0833	3.0031 (0.3088)	0.0953	4.0013 (0.3722)	0.1384	1.9978 (0.3676)	0.1350
		100	1.9930 (0.1946)	0.0379	1.0054 (0.1919)	0.0368	0.4935 (0.1917)	0.0367	3.0049 (0.1929)	0.0372	3.9966 (0.1677)	0.0281	2.0063 (0.1676)	0.0281
		250	2.0017 (0.1085)	0.0118	0.9981 (0.1099)	0.0121	0.4985 (0.1066)	0.0114	3.0012 (0.1085)	0.0118	3.9916 (0.1126)	0.0127	2.0096 (0.1126)	0.0127
	Moderate linearity	10	1.8465 (1.8306)	3.3714	1.0939 (1.4460)	2.0977	1.1512 (1.6142)	3.0271	2.9423 (2.3770)	5.6476	3.6354 (2.6243)	7.0131	2.0653 (2.2439)	5.0343
		30	1.9908 (1.3001)	1.6886	1.0672 (0.9941)	0.9918	0.7152 (0.8651)	0.7939	3.0830 (1.4355)	2.0655	3.8480 (1.5026)	2.2788	1.8965 (1.3388)	1.8012
		100	1.9950 (0.7956)	0.6324	0.9698 (0.6792)	0.4618	0.6310 (0.6088)	0.3874	2.9919 (0.8441)	0.7118	3.9619 (0.6846)	0.4697	1.9652 (0.6838)	0.4684
		250	1.9929 (0.4226)	0.1784	0.9978 (0.4209)	0.1770	0.5172 (0.3852)	0.1485	3.0119 (0.4554)	0.2073	4.0091 (0.4484)	0.2010	1.9725 (0.4529)	0.2057
	Low linearity	10	2.2697 (3.0926)	9.6273	1.7127 (2.7583)	8.1086	1.5774 (2.6388)	8.1173	2.6810 (3.3493)	11.3084	3.3791 (3.8545)	15.2275	2.1107 (3.0617)	9.3766
		30	2.2380 (2.3280)	5.4710	1.3373 (1.7476)	3.1650	1.0890 (1.5753)	2.8261	2.9227 (2.5249)	6.3747	3.5889 (2.5818)	6.8282	1.9409 (2.1291)	4.5320
		100	2.0963 (1.4435)	2.0907	1.0337 (1.1569)	1.3383	0.8978 (1.0614)	1.2838	2.8695 (1.6152)	2.6234	3.8681 (1.4090)	2.0008	1.8298 (1.2690)	1.6376
		250	1.9795 (0.8919)	0.7952	1.0138 (0.7614)	0.5793	0.6147 (0.6243)	0.4024	3.0427 (0.9133)	0.8352	3.9274 (0.8772)	0.7740	1.9462 (0.8834)	0.7825
Normal	High linearity	10	1.9335 (1.1022)	1.2180	1.0796 (1.0699)	1.1499	0.8044 (0.9239)	0.9454	2.7165 (0.9927)	1.0648	3.7319 (1.7844)	3.2528	2.2395 (1.7401)	3.0822
		30	1.8960 (0.9010)	0.8219	1.1043 (0.8914)	0.8046	0.6619 (0.6729)	0.4785	2.8744 (0.7193)	0.5326	3.9646 (0.9385)	0.8812	2.0116 (0.9385)	0.8800
		100	2.0191 (0.4581)	0.2100	0.9792 (0.4589)	0.2108	0.5275 (0.4039)	0.1638	2.9795 (0.4103)	0.1686	4.0082 (0.4896)	0.2396	1.9881 (0.4889)	0.2389
		250	2.0111 (0.3065)	0.0940	0.9885 (0.3075)	0.0946	0.5020 (0.2882)	0.0830	3.0003 (0.2909)	0.0845	3.9960 (0.3062)	0.0937	2.0024 (0.3065)	0.0939
	Moderate linearity	10	1.7939 (1.6873)	2.8866	1.3559 (1.5790)	2.6173	1.4072 (1.6526)	3.5514	2.1019 (1.8264)	4.1390	3.4469 (3.0061)	9.3338	2.6448 (2.8313)	8.4238
		30	1.7653 (1.4340)	2.1095	1.3068 (1.3996)	2.0509	1.2884 (1.5964)	3.1675	2.4572 (1.7907)	3.4979	3.4883 (2.3347)	5.7071	2.3428 (2.2566)	5.2046
		100	1.7294 (1.2176)	1.5544	1.3085 (1.1984)	1.5300	0.9105 (1.1299)	1.4440	2.6305 (1.2046)	1.5862	3.8351 (1.7039)	2.9276	2.1196 (1.6542)	2.7481
		250	1.9704 (0.9687)	0.9384	1.0424 (0.9336)	0.8724	0.7582 (0.8749)	0.8313	2.7863 (0.9319)	0.9132	3.9787 (1.1849)	1.4030	1.9858 (1.1754)	1.3804
	Low linearity	10	1.8112 (2.1713)	4.7455	1.6030 (2.0915)	4.7334	1.5201 (1.9958)	5.0199	1.9307 (2.1994)	5.9760	3.3157 (3.5242)	12.8760	2.9620 (8.6705)	12.3616
		30	1.6961 (1.6155)	2.8155	1.4765 (2.0717)	2.8343	1.7566 (2.2446)	5.8665	2.1536 (2.8527)	5.7495	3.3980 (2.6295)	8.4922	2.3642 (2.6295)	7.0399
		100	1.7324 (1.4545)	2.1852	1.3631 (1.4275)	2.1677	1.3286 (1.5222)	3.0012	2.2413 (1.6568)	3.3178	3.5421 (2.4205)	6.0628	2.4196 (2.3738)	5.8052
		250	1.8744 (1.2642)	1.6123	1.1821 (1.2355)	1.5581	1.1402 (1.3295)	2.1756	2.4338 (1.4268)	2.3542	3.7001 (1.9254)	3.7933	2.2343 (1.8837)	3.5997
Log-Normal	High linearity	10	1.9995 (0.2592)	0.0671	0.9992 (0.1305)	0.0170	0.5030 (0.3632)	0.1318	2.9993 (0.2611)	0.0681	3.9940 (0.6491)	0.4209	2.0065 (0.3573)	0.1276
		30	2.0190 (0.2334)	0.0548	0.9975 (0.1148)	0.0132	0.4902 (0.1843)	0.0340	2.9998 (0.1000)	0.0100	3.9916 (0.2187)	0.0479	2.0035 (0.0908)	0.0083
		100	2.0008 (0.0991)	0.0098	1.0003 (0.0511)	0.0026	0.4972 (0.0904)	0.0082	3.0003 (0.0454)	0.0021	4.0015 (0.1456)	0.0212	1.9985 (0.0697)	0.0049
		250	2.0032 (0.0611)	0.0037	0.9991 (0.0296)	8.7710E-4	0.4984 (0.0686)	0.0047	3.0002 (0.0329)	0.0011	3.9988 (0.0698)	0.0049	2.0008 (0.0271)	7.3328E-4
	Moderate linearity	10	2.0342 (0.9884)	0.9771	1.0111 (0.5102)	0.2601	0.8533 (1.0304)	1.1855	2.9503 (0.9952)	0.9919	3.7173 (2.3047)	5.3861	1.9883 (1.3508)	1.8230
		30	1.9466 (0.8379)	0.7043	0.9996 (0.4538)	0.2057	0.6020 (0.5757)	0.3425	2.9935 (0.4076)	0.1660	3.9811 (0.8417)	0.1295	1.9913 (0.3599)	0.1295
		100	1.9948 (0.3900)	0.1520	1.0021 (0.2065)	0.0426	0.5003 (0.3283)	0.1077	3.0020 (0.1746)	0.0304	4.0101 (0.5487)	0.3008	1.9911 (0.2764)	0.0764
		250	1.9978 (0.2461)	0.0605	1.0030 (0.1237)	0.0153	0.5035 (0.2600)	0.0675	2.9953 (0.1282)	0.0164	4.0054 (0.2809)	0.0112	1.9997 (0.1057)	0.0112
	Low linearity	10	2.0087 (1.5091)	2.2751	0.9506 (0.8067)	0.6526	1.1338 (1.5730)	2.8736	2.8054 (1.6660)	2.8736	3.7443 (3.3254)	11.1125	2.2112 (2.0056)	4.0632
		30	1.8423 (1.4035)	1.9927	1.0917 (0.8283)	0.6938	0.8012 (0.9516)	0.9954	2.9506 (0.7837)	0.6160	3.8726 (1.7012)	2.9073	1.9532 (0.7371)	0.5450
		100	1.9162 (0.7655)	0.5924	1.0074 (0.3853)	0.1484	0.5987 (0.5757)	0.3409	2.9924 (0.3517)	0.1236	4.0254 (1.1206)	1.2552	1.9845 (0.5615)	0.3152
		250	1.9824 (0.4878)	0.2380	0.9982 (0.2319)	0.0537	0.5439 (0.4616)	0.2147	3.0002 (0.2520)	0.0634	3.9881 (0.5582)	0.3115	2.0003 (0.2086)	0.0435
Mixture of distributions	High linearity	10	1.9363 (0.3451)	0.1230	1.0299 (0.3530)	0.1254	0.9363 (1.1184)	1.4399	2.6632 (1.2904)	1.7769	3.9781 (0.4013)	0.1614	1.9683 (0.3069)	0.0951
		30	1.9915 (0.1996)	0.0399	1.0007 (0.1876)	0.0352	0.5537 (0.4041)	0.1660	3.0010 (0.3436)	0.1179	3.9488 (0.4385)	0.1947	2.0029 (0.3321)	0.1102
		100	1.9958 (0.1928)	0.0372	1.0060 (0.1677)	0.0281	0.5019 (0.1780)	0.0317	2.9964 (0.1432)	0.0205	4.0035 (0.1889)	0.0356	1.9990 (0.1658)	0.0275
		250	1.9841 (0.2830)	0.0803	1.0180 (0.2711)	0.0738	0.4995 (0.1109)	0.0123	2.9992 (0.0923)	0.0085	4.0015 (0.1080)	0.0116	1.9981 (0.0887)	0.0079
	Moderate linearity	10	1.8474 (1.2568)	1.6012	1.1144 (0.9269)	0.8714	1.6007 (1.8294)	4.5548	2.4494 (2.3034)	5.6036	3.8003 (1.6284)	2.6888	1.7961 (1.1116)	1.2760
		30	1.8712 (0.7857)	0.6332	0.9493 (0.6842)	0.4703	0.9697 (1.1574)	1.5589	3.2215 (1.4490)	2.1466	3.6324 (1.4992)	2.3805	1.8670 (1.2896)	1.6790
		100	1.9627 (0.7796)	0.6085	0.9737 (0.6110)	0.3737	0.6024 (0.5535)	0.3165	3.0458 (0.6217)	0.3882	3.9526 (0.7588)	0.5774	1.9699 (0.6225)	0.3880
		250	1.9224 (1.0672)	1.1437	1.1499 (0.9445)	0.9137	0.5197 (0.4060)	0.1650	3.0055 (0.3868)	0.1495	3.9735 (0.4215)	0.1782	1.9479 (0.3665)	0.1369
	Low linearity	10	2.1516 (2.0561)	4.2464	1.3251 (1.6699)	2.8915	1.5551 (2.2007)	5.9513	2.2149 (2.6162)	7.4541	3.7847 (2.9630)	8.8171	1.7467 (1.9410)	3.8280
		30	1.8443 (1.4731)	2.1922	1.0391 (1.1439)	1.3087	1.4430 (1.9353)	4.6306	3.0519 (2.3644)	5.5875	3.1173 (2.4727)	6.8876	2.1166 (2.2135)	4.9081
		100	1.9390 (1.4486)	2.0999	1.0298 (1.0009)	1.0017	0.7762 (0.9707)	1.0176	3.1882 (1.2535)	1.6050	3.8126 (1.4490)	2.1326	1.7445 (1.1104)	1.2970
		250	1.9957 (1.7411)	3.0284	1.4922 (1.5705)	2.7063	0.5926 (0.6582)	0.4414	3.0346 (0.8909)	0.7941	3.7326 (0.7978)	0.7073	1.7582 (0.7517)	0.6230

Table 10: Results of the *DSD Model* $\Psi^{-1}_{\widehat{Y}(j)}(t) = -1 + 2\Psi^{-1}_{X_1(j)}(t) - 1\Psi^{-1}_{X_1(j)}(1-t) + 0.5\Psi^{-1}_{X_2(j)}(t) - 3\Psi^{-1}_{X_2(j)}(1-t) + 4\Psi^{-1}_{X_3(j)}(t) - 2\Psi^{-1}_{X_3(j)}(1-t)$ in different conditions.

Distributions in variables $X_1; X_2; X_3$	*Degree of linearity*	n	*Parameter estimated*		*Goodness-of-fit measures*			
			$\overline{\widehat{\gamma}}\,(s)$	MSE(γ)	$\overline{\Omega}\,(s)$	$\overline{RMSE}_M\,(s)$	$\overline{RMSE}_L\,(s)$	$\overline{RMSE}_U\,(s)$
Uniform	High linearity	10	$-0.9954\,(0.4265)$	0.1817	$0.9802\,(0.0086)$	$1.0224\,(0.2335)$	$0.8774\,(0.2080)$	$0.8810\,(0.2104$
		30	$-1.0036\,(0.2676)$	0.0711	$0.9713\,(0.0054)$	$1.3036\,(0.1275)$	$1.1242\,(0.1248)$	$1.1303\,(0.1260$
		100	$-1.0029\,(0.1462)$	0.0214	$0.9682\,(0.0030)$	$1.4041\,(0.0690)$	$1.2229\,(0.0701)$	$1.2293\,(0.0713$
		250	$-0.9981\,(0.0900)$	0.0081	$0.9673\,(0.0019)$	$1.3925\,(0.0411)$	$1.2150\,(0.0415)$	$1.2216\,(0.0420$
	Moderate linearity	10	$-0.8303\,(1.6879)$	2.8749	$0.7755\,(0.0760)$	$4.1952\,(0.8943)$	$3.6459\,(0.7989)$	$3.6639\,(0.8095)$
		30	$-1.0236\,(1.0221)$	1.0441	$0.6907\,(0.0445)$	$5.2222\,(0.5082)$	$4.5245\,(0.4971)$	$4.5509\,(0.5021)$
		100	$-1.0353\,(0.5694)$	0.3252	$0.6579\,(0.0241)$	$5.6366\,(0.2791)$	$4.9108\,(0.2867)$	$4.9383\,(0.2910)$
		250	$-0.9939\,(0.3537)$	0.1250	$0.6502\,(0.0146)$	$5.5715\,(0.1699)$	$4.8617\,(0.1735)$	$4.8881\,(0.1762)$
	Low linearity	10	$-1.1547\,(3.3959)$	11.5444	$0.5209\,(0.1136)$	$8.5895\,(1.8086)$	$7.5357\,(1.6751)$	$7.5766\,(1.6945)$
		30	$-0.9447\,(2.0460)$	4.1914	$0.3841\,(0.0608)$	$10.5360\,(1.0068)$	$9.1760\,(0.9728)$	$9.2280\,(0.9887)$
		100	$-1.0769\,(1.1387)$	1.3014	$0.3342\,(0.0284)$	$11.2831\,(0.5547)$	$9.8354\,(0.5644)$	$9.8882\,(1.1387)$
		250	$-1.0020\,(0.7064)$	0.4984	$0.3205\,(0.0158)$	$11.1654\,(0.3276)$	$9.7469\,(0.3322)$	$9.7996\,(0.3368)$
Normal	High linearity	10	$-0.9551\,(2.6938)$	7.2511	$0.9807\,(0.0076)$	$2.3881\,(0.4903)$	$2.1362\,(0.4250)$	$2.1787\,(0.4423)$
		30	$-1.0019\,(1.4297)$	2.0419	$0.9760\,(0.0047)$	$2.7536\,(0.2763)$	$2.4042\,(0.2452)$	$2.4687\,(0.2557)$
		100	$-1.0583\,(0.8375)$	0.7040	$0.9734\,(0.0028)$	$2.8654\,(0.1526)$	$2.4964\,(0.1386)$	$2.5653\,(0.1487)$
		250	$-1.0062\,(0.5086)$	0.2585	$0.9727\,(0.0017)$	$2.9668\,(0.0955)$	$2.5845\,(0.0884)$	$2.6554\,(0.0939)$
	Moderate linearity	10	$-0.9392\,(5.6528)$	31.9262	$0.7541\,(0.0708)$	$10.1428\,(1.9894)$	$9.0345\,(1.7548)$	$9.2209\,(1.8396)$
		30	$-0.8546\,(3.8886)$	15.1275	$0.7198\,(0.0417)$	$11.1323\,(1.1321)$	$9.7282\,(1.0332)$	$9.9828\,(1.1018)$
		100	$-0.9645\,(2.887)$	6.6959	$0.6997\,(0.0238)$	$11.4256\,(0.6276)$	$9.9600\,(0.5695)$	$10.2324\,(0.6053)$
		250	$-1.1650\,(1.8498)$	3.4455	$0.6913\,(0.0146)$	$11.8738\,(0.3993)$	$10.3490\,(0.3623)$	$10.6280\,(0.3869)$
	Low linearity	10	$-0.5297\,(8.6705)$	75.3237	$0.4439\,(0.0969)$	$21.0218\,(3.9501)$	$18.6861\,(3.5548)$	$19.0921\,(3.7412)$
		30	$-0.9447\,(5.9433)$	35.2910	$0.3919\,(0.0506)$	$22.7387\,(2.1426)$	$19.8675\,(1.9448)$	$20.3988\,(2.0596)$
		100	$-1.0364\,(3.9277)$	15.4129	$0.3722\,(0.0276)$	$22.9152\,(1.2349)$	$19.9788\,(1.1392)$	$20.5125\,(1.2073)$
		250	$-1.1309\,(3.0232)$	9.1475	$0.3621\,(0.0170)$	$23.7122\,(0.7765)$	$20.6789\,(0.7120)$	$21.2253\,(0.7660)$
Log-Normal	High linearity	10	$-0.9704\,(1.8372)$	7.2543	$0.9804\,(0.0076)$	$5.0253\,(1.0173)$	$4.3769\,(0.9066)$	$4.8584\,(0.9644)$
		30	$-1.0003\,(0.9822)$	4.9651	$0.9769\,(0.0049)$	$5.1463\,(0.5603)$	$4.3645\,(0.4610)$	$4.8026\,(0.5021)$
		100	$-0.9751\,(0.5787)$	4.2904	$0.9751\,(0.0029)$	$5.4867\,(0.3234)$	$4.6229\,(0.2622)$	$5.0578\,(0.2858)$
		250	$-1.0002\,(0.3863)$	4.1499	$0.9764\,(0.0017)$	$6.1185\,(0.2209)$	$5.1157\,(0.1761)$	$5.7526\,(0.1942)$
	Moderate linearity	10	$-1.1808\,(7.0232)$	54.0323	$0.7619\,(0.0672)$	$20.4618\,(3.8014)$	$17.8046\,(3.3664)$	$19.7133\,(3.6410)$
		30	$-1.0620\,(3.7799)$	18.5250	$0.7276\,(0.0453)$	$20.6888\,(2.2543)$	$17.5517\,(1.8580)$	$19.2404\,(2.0258)$
		100	$-0.9793\,(2.2593)$	9.0170	$0.7117\,(0.0256)$	$21.9200\,(1.2722)$	$18.4948\,(1.0062)$	$20.2220\,(1.1156)$
		250	$-1.0131\,(1.6101)$	6.6424	$0.7225\,(0.0155)$	$24.4507\,(0.8628)$	$20.4471\,(0.7119)$	$22.9771\,(0.7844)$
	Low linearity	10	$-2.3100\,(14.3309)$	216.1254	$0.4693\,(0.1061)$	$40.9232\,(7.3866)$	$35.4632\,(6.4742)$	$39.4028\,(6.9347)$
		30	$-1.0768\,(8.1259)$	70.2765	$0.4064\,(0.0657)$	$41.5803\,(4.4757)$	$35.5610\,(3.6308)$	$38.5610\,(3.9304)$
		100	$-1.2582\,(4.5557)$	25.8335	$0.3862\,(0.0359)$	$43.7592\,(2.6744)$	$36.9316\,(2.1379)$	$40.3388\,(2.3526)$
		250	$-1.1111\,(3.1352)$	14.2765	$0.3950\,(0.0237)$	$48.9454\,(1.7670)$	$40.9188\,(1.4259)$	$45.9808\,(1.5573)$
Mixture of distributions	High linearity	10	$-1.7320\,(2.2687)$	5.6778	$0.9794\,(0.0079)$	$2.1996\,(0.4428)$	$1.99556\,(0.3972)$	$2.0054\,(0.4130)$
		30	$-0.9887\,(0.7195)$	0.5173	$0.9732\,(0.0051)$	$2.5414\,(0.2522)$	$2.2412\,(0.2359)$	$2.2870\,(0.2432)$
		100	$-1.0048\,(0.3215)$	0.1033	$0.9735\,(0.0029)$	$2.7786\,(0.1539)$	$2.4032\,(0.1365)$	$2.5040\,(0.1450)$
		250	$-0.9672\,(0.5479)$	0.3009	$0.9702\,(0.0019)$	$2.8893\,(0.0936)$	$2.5135\,(0.0862)$	$2.5828\,(0.0903)$
	Moderate linearity	10	$-2.5969\,(4.7119)$	24.7298	$0.7577\,(0.0741)$	$8.9627\,(1.8086)$	$8.0426\,(1.6017)$	$8.1954\,(1.6694)$
		30	$-0.9733\,(2.7122)$	7.3494	$0.6999\,(0.0418)$	$10.1960\,(0.9850)$	$9.0024\,(0.9229)$	$9.1889\,(0.9547)$
		100	$-0.9637\,(1.3067)$	1.7071	$0.6998\,(0.0249)$	$11.0752\,(0.6139)$	$9.5849\,(0.5434)$	$9.9884\,(0.5684)$
		250	$-0.8069\,(2.0196)$	4.1120	$0.6717\,(0.0149)$	$11.5622\,(0.3730)$	$10.0519\,(0.3491)$	$10.3373\,(0.3676)$
	Low linearity	10	$-2.9719\,(7.4423)$	59.2206	$0.4813\,(0.1093)$	$18.1495\,(3.4327)$	$16.2819\,(3.0929)$	$16.6343\,(3.2040)$
		30	$-0.8047\,(5.4384)$	29.5848	$0.3888\,(0.0530)$	$20.2975\,(2.0511)$	$17.8591\,(1.9131)$	$18.2629\,(1.9743)$
		100	$-1.1448\,(2.6361)$	6.9628	$0.3728\,(0.0322)$	$22.1942\,(1.2382)$	$19.1900\,(1.1028)$	$19.9925\,(1.1661)$
		250	$-0.4857\,(3.4173)$	11.9304	$0.3414\,(0.0170)$	$23.1095\,(0.7457)$	$20.0925\,(0.6992)$	$20.6545\,(0.7294)$

Table 11: Results of the *DSD Model* $\Psi^{-1}_{\widehat{Y}(j)}(t) = -1 + 2\Psi^{-1}_{X_1(j)}(t) - 1\Psi^{-1}_{X_1(j)}(1-t) + 0.5\Psi^{-1}_{X_2(j)}(t) - 3\Psi^{-1}_{X_2(j)}(1-t) + 4\Psi^{-1}_{X_3(j)}(t) - 2\Psi^{-1}_{X_3(j)}(1-t)$ in different conditions (continuation of the *Table 10*).

Appendix D: Observed and predicted histograms of the experiments presented in *Subsection 4.2.1*.

Histogram-valued variable Hematocrit in relation with the histogram-valued variable Hemoglobin

In the example in *Subsection 4.2.1* we performed a comparative study of the *DSD Model* with other existing models. The results of the application of the models proposed by Billard-Diday [5] and Verde-Irpino [23] to the data of this example may be found in *Table 12*.

DSD Model	$\Psi^{-1}_{\widehat{Y}(j)}(t) = -1.953 + 3.5598\Psi^{-1}_{X(j)}(t) - 0.4128\Psi^{-1}_{X(j)}(1-t)$
Billard-Diday Model	$\widehat{\underline{I}}_{Y(j)_i} = -2.16 + 3.16\underline{I}_{X(j)_i} \qquad \widehat{\overline{I}}_{Y(j)_i} = -2.16 + 3.16\overline{I}_{X(j)_i}$
Verde-Irpino Model	$\Psi^{-1}_{\widehat{Y}(j)}(t) = -2.157 + 3.161\overline{X}(j) + 3.918\left(\Psi^{-1}_{X(j)}(t) - \overline{X}(j)\right)$

Table 12: Linear regression models applied to the data in *Table 4*.

In *Table 13*, in the white rows we have the observed histograms of each observation of the histogram-valued variable Y, in the light grey rows the histograms $H_{\widehat{Y}_{DSD}(j)}$ predicted using the *DSD Model*, in the grey rows the histograms $H_{\widehat{Y}_{BD}(j)}$ predicted using the model proposed by Billard and Diday [5] and in the dark grey the histograms $H_{\widehat{Y}_{VI}(j)}$ predicting with the Verde and Irpino [23].

Obs.	Distributions of the values of hematocrit
$H_{Y(1)}$	{[33.29; 35.41[, 0.3; [35.41; 36.11[, 0.1; [36.11; 36.82[, 0.1; [36.82; 37.52[, 0.1; [37.52; 38.04[, 0.1; [38.04; 39.61] , 0.3}
$H_{\widehat{Y}_{DSD}(1)}$	{[33.84; 35.70[, 0.3; [35.70; 36.32[, 0.1; [36.32; 36.73[, 0.1; [36.73; 37.13[, 0.1; [37.13; 37.56[, 0.1; [37.56; 38.85] , 0.3}
$H_{\widehat{Y}_{BD}(1)}$	{[34.33; 35.87[, 0.3; [35.87; 36.38[, 0.1; [36.38; 36.70[, 0.1; [36.70; 37.02[, 0.1; [37.02; 37.35[, 0.1; [37.35; 38.31] , 0.3}
$H_{\widehat{Y}_{VI}(1)}$	{[33.79; 35.7[, 0.3; [35.7; 36.34[, 0.1; [36.34; 36.73[, 0.1; [36.73; 37.13[, 0.1; [37.13; 37.53[, 0.1; [37.53; 38.73] , 0.3}
$H_{Y(2)}$	{[36.69; 39.11[, 0.3; [39.11; 39.97[, 0.1; [39.97; 40.83[, 0.1; [40.83; 41.69[, 0.1; [41.69; 42.54[, 0.1; [42.54; 45.12] , 0.3}
$H_{\widehat{Y}_{DSD}(2)}$	{[35.16; 38.04[, 0.3; [38.04; 39.00[, 0.1; [39.00; 39.96[, 0.1; [39.96; 40.67[, 0.1; [40.67; 41.38[, 0.1; [41.38; 43.51] , 0.3}
$H_{\widehat{Y}_{BD}(2)}$	{[36.00; 38.37[, 0.3; [38.37; 39.16[, 0.1; [39.16; 39.95[, 0.1; [39.95; 40.49[, 0.1; [40.49; 41.03[, 0.1; [41.03; 42.64] , 0.3}
$H_{\widehat{Y}_{VI}(2)}$	{[35.13; 38.06[, 0.3; [38.06; 39.04[, 0.1; [39.04; 40.02[, 0.1; [40.02; 40.69[, 0.1; [40.69; 41.36[, 0.1; [41.36; 43.35] , 0.3}
$H_{Y(3)}$	{[36.69; 40.26[, 0.3; [40.26; 41.45[, 0.1; [41.45; 42.64[, 0.1; [42.64; 43.85[, 0.1; [43.85; 45.06[, 0.1; [45.06; 48.68] , 0.3}
$H_{\widehat{Y}_{DSD}(3)}$	{[35.45; 42.27[, 0.3; [42.27; 43.38[, 0.1; [43.38; 44.50[, 0.1; [44.50; 45.61[, 0.1; [45.61; 46.72[, 0.1; [46.72; 50.46] , 0.3}
$H_{\widehat{Y}_{BD}(3)}$	{[36.98; 42.74[, 0.3; [42.74; 43.62[, 0.1; [43.62; 44.51[, 0.1; [44.51; 45.39[, 0.1; [45.39; 46.28[, 0.1; [46.28; 48.93] , 0.3}
$H_{\widehat{Y}_{VI}(3)}$	{[35.2942.42[, 0.3; [42.42; 43.51[, 0.1; [43.51; 44.61[, 0.1; [44.61; 45.71[, 0.1; [45.71; 46.80[, 0.1; [46.80; 50.1] , 0.3}
$H_{Y(4)}$	{[36.38; 39.75[, 0.3; [39.75; 40.87[, 0.1; [40.87; 41.96[, 0.1; [41.96; 43.05[, 0.1; [43.05; 44.14[, 0.1; [44.14; 47.41] , 0.3}
$H_{\widehat{Y}_{DSD}(4)}$	{[35.80; 40.08[, 0.3; [40.08; 41.50[, 0.1; [41.50; 42.92[, 0.1; [42.92; 43.81[, 0.1; [43.81; 44.70[, 0.1; [44.70; 47.37] , 0.3}
$H_{\widehat{Y}_{BD}(4)}$	{[36.98; 40.55[, 0.3; [40.55; 41.74[, 0.1; [41.74; 42.93[, 0.1; [42.93; 43.58[, 0.1; [43.58; 44.23[, 0.1; [44.23; 46.18] , 0.3}
$H_{\widehat{Y}_{VI}(4)}$	{[35.7140.13[, 0.3; [40.13; 41.61[, 0.1; [41.61; 43.08[, 0.1; [43.08; 43.89[, 0.1; [43.89; 44.69[, 0.1; [44.69; 47.12] , 0.3}
$H_{Y(5)}$	{[39.19; 42.69[, 0.3; [42.69; 43.86[, 0.1; [43.86; 45.03[, 0.1; [45.03; 46.19[, 0.1; [46.19; 47.36[, 0.1; [47.36; 50.86] , 0.3}
$H_{\widehat{Y}_{DSD}(5)}$	{[39.68; 42.52[, 0.3; [42.52; 43.64[, 0.1; [43.64; 44.75[, 0.1; [44.75; 45.86[, 0.1; [45.86; 46.97[, 0.1; [46.97; 50.25] , 0.3}
$H_{\widehat{Y}_{BD}(5)}$	{[40.78; 42.99[, 0.3; [42.99; 43.87[, 0.1; [43.87; 44.76[, 0.1; [44.76; 45.64[, 0.1; [45.64; 46.53[, 0.1; [46.53; 49.19] , 0.3}
$H_{\widehat{Y}_{VI}(5)}$	{[39.842.54[, 0.3; [42.54; 43.64[, 0.1; [43.64; 44.74[, 0.1; [44.74; 45.83[, 0.1; [45.83; 46.93[, 0.1; [46.93; 50.22] , 0.3}
$H_{Y(6)}$	{[39.70; 43.17[, 0.3; [43.17; 44.32[, 0.1; [44.32; 44.81[, 0.1; [44.81; 45.29[, 0.1; [45.29; 45.78[, 0.1; [45.78; 47.24] , 0.3}
$H_{\widehat{Y}_{DSD}(6)}$	{[40.93; 42.92[, 0.3; [42.92; 43.58[, 0.1; [43.58; 44.04[, 0.1; [44.04; 44.51[, 0.1; [44.51; 44.99[, 0.1; [44.99; 46.45] , 0.3}
$H_{\widehat{Y}_{BD}(6)}$	{[41.50; 43.14[, 0.3; [43.14; 43.68[, 0.1; [43.68; 44.05[, 0.1; [44.05; 44.42[, 0.1; [44.42; 44.79[, 0.1; [44.79; 45.90] , 0.3}
$H_{\widehat{Y}_{VI}(6)}$	{[40.9242.95[, 0.3; [42.95; 43.62[, 0.1; [43.62; 44.08[, 0.1; [44.08; 44.54[, 0.1; [44.54; 44.99[, 0.1; [44.99; 46.47] , 0.3}
$H_{Y(7)}$	{[41.56; 44.11[, 0.3; [44.11; 44.95[, 0.1; [44.95; 45.80[, 0.1; [45.80; 46.65[, 0.1; [46.65; 47.19[, 0.1; [47.19; 48.81] , 0.3}
$H_{\widehat{Y}_{DSD}(7)}$	{[42.67; 43.86[, 0.3; [43.86; 44.26[, 0.1; [44.26; 44.65[, 0.1; [44.65; 45.22[, 0.1; [45.22; 45.78[, 0.1; [45.78; 47.48] , 0.3}
$H_{\widehat{Y}_{BD}(7)}$	{[43.18; 44.07[, 0.3; [44.07; 44.37[, 0.1; [44.37; 44.66[, 0.1; [44.66; 45.13[, 0.1; [45.13; 45.60[, 0.1; [45.60; 47.00] , 0.3}
$H_{\widehat{Y}_{VI}(7)}$	{[42.7643.87[, 0.3; [43.87; 44.24[, 0.1; [44.24; 44.61[, 0.1; [44.61; 45.19[, 0.1; [45.19; 45.77[, 0.1; [45.77; 47.51] , 0.3}
$H_{Y(8)}$	{[38.4; 40.34[, 0.3; [40.34; 40.99[, 0.1; [40.99; 41.64[, 0.1; [41.64; 42.28[, 0.1; [42.28; 42.93[, 0.1; [42.93; 45.22] , 0.3}
$H_{\widehat{Y}_{DSD}(8)}$	{[39.26; 40.74[, 0.3; [40.74; 41.24[, 0.1; [41.24; 41.72[, 0.1; [41.72; 42.20[, 0.1; [42.20; 42.79[, 0.1; [42.79; 44.54] , 0.3}
$H_{\widehat{Y}_{BD}(8)}$	{[39.80; 40.95[, 0.3; [40.95; 41.33[, 0.1; [41.33; 41.72[, 0.1; [41.72; 42.10[, 0.1; [42.10; 42.58[, 0.1; [42.58; 44.00] , 0.3}
$H_{\widehat{Y}_{VI}(8)}$	{[39.3140.74[, 0.3; [40.74; 41.22[, 0.1; [41.22; 41.7[, 0.1; [41.7; 42.17[, 0.1; [42.17; 42.76[, 0.1; [42.76; 44.52] , 0.3}
$H_{Y(9)}$	{[28.83; 32.86[, 0.3; [32.86; 34.21[, 0.1; [34.21; 35.55[, 0.1; [35.55; 36.84[, 0.1; [36.84; 38.12[, 0.1; [38.12; 41.98] , 0.3}
$H_{\widehat{Y}_{DSD}(9)}$	{[27.66; 33.54[, 0.3; [33.54; 35.50[, 0.1; [35.50; 36.70[, 0.1; [36.70; 37.91[, 0.1; [37.91; 39.20[, 0.1; [39.20; 43.08] , 0.3}
$H_{\widehat{Y}_{BD}(9)}$	{[29.20; 34.09[, 0.3; [34.09; 35.72[, 0.1; [35.72; 36.68[, 0.1; [36.68; 37.63[, 0.1; [37.63; 38.59[, 0.1; [38.59; 41.47] , 0.3}
$H_{\widehat{Y}_{VI}(9)}$	{[27.5433.59[, 0.3; [33.59; 35.61[, 0.1; [35.61; 36.8[, 0.1; [36.8; 37.9[, 0.1; [37.9; 39.18[, 0.1; [38.18; 42.74] , 0.3}
$H_{Y(10)}$	{[44.48; 46.90[, 0.3; [46.90; 47.70[, 0.1; [47.70; 48.51[, 0.1; [48.51; 49.31[, 0.1; [49.31; 50.12[, 0.1; [50.12; 52.53] , 0.3}
$H_{\widehat{Y}_{DSD}(10)}$	{[45.85; 47.48[, 0.3; [47.48; 48.03[, 0.1; [48.03; 48.58[, 0.1; [48.58; 49.13[, 0.1; [49.13; 49.68[, 0.1; [49.68; 51.33] , 0.3}
$H_{\widehat{Y}_{BD}(10)}$	{[46.43; 47.73[, 0.3; [47.73; 48.17[, 0.1; [48.17; 48.61[, 0.1; [48.61; 49.05[, 0.1; [49.05; 49.48[, 0.1; [49.48; 50.80] , 0.3}
$H_{\widehat{Y}_{VI}(10)}$	{[45.9147.51[, 0.3; [47.51; 48.06[, 0.1; [48.06; 48.6[, 0.1; [48.6; 49.14[, 0.1; [49.14; 49.68[, 0.1; [49.68; 51.31] , 0.3}

Table 13: Observed and predicted histograms (using three different methods) of the Hematocrit values for the data in *Table 4*.